Music Genre/Tempo based Aroma Release System

And

Linear and Non-Linear Global Feature based Classification of Emotional Speech

Report submitted to:

Indian Institute of Technology, Kharagpur

For the award of degree

Of

Masters of Technology in

Instrumentation and Signal Processing

Electrical Engineering

by

Gursewak Singh Sidhu

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY,

KHARAGPUR June, 2016

CERTIFICATE

This is to certify that Dissertation Report entitled, "**Music Genre/Tempo based Aroma Release System && Linear and Non-Linear Global features based classification of Emotional Speech**", submitted by **Mr Gursewak Singh Sidhu** to **Indian Institute of Technology**, **Kharagpur** is a record of bonafide project work carried out by him under my supervision and guidance in the Department of Electrical Engineering.

_____ _____

Prof. Aurobinda Routray **External Examiner**

DECELERATION

I certify that

a. The work contained in this report is original and has been done by me under the guidance of my supervisor, Prof. Aurobinda Routray.
b. The work has not been submitted to any other Institute for any degree or diploma.
c. I have followed the guidelines provided by the Institute in preparing the report.
d. I have conformed to the norms and guidelines given in the Ethical Code of conduct of the Institute.
e. Wherever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of report and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

_____

Gursewak Singh Sidhu

ACKNOWLEDGEMENT

I take the opportunity to express my reverence to my supervisor Prof. Aurobinda Routray for his guidance, inspiration and innovative technical discussion during the course of this project. His perpetual energy and enthusiasm in research has motivated us. In addition, he was always accessible and willing to help in this project.

I would also like to thank Dr. Rajlakshmi Guha (Counsellor, Student Counselling Centre) and Dr. Rashmi Mukherjee, (Post-Doctoral Scientist, Indian Institute of Technology, Kharagpur Metabolomics, Human Monitoring Devices, Quantitative Microscopy, Biomarker discovery, Cognitive Neuroscience) for her support, who always found time between their busy research schedules to help me when stuck at any problem.

I would also like to thank Suvodip Chakraborty and Supriyo Chakraborty (Research Scholars, Electrical Engineering Department), for help in conducting Trier Social Stress Experiment. I would also like to extend my gratitude to all the professors for their contribution in my studies and research work. They have been great source of inspiration for me.

Last but not the least I would like to thank my parents, who taught the value of hard work by their own example.

Gursewak Singh Sidhu,
Roll no.: **11EE35013**
Department of Electrical Engineering
IIT Kharagpur

Part A:
Music Genre/Tempo based Aroma Release System

MOTIVATION

Moving from the time of mute movies to colour cinema and now with 3D and 4D TV and Cinema, a lot has happened in cinema industry, but when it comes to listening to music, a little has changed with respect to experience of user. Music has been associated with every known culture of society. Aromatherapy is a well-established field which has physiological and psychological effects on humans. In this project we try to bring both together to enhance the pleasure of subject.

ABSTRACT

In this project we try to classify the music being played in real time using a standalone embedded system and releases a specific aroma pertaining to the genre to increase the calming, soothing effect on brain and enhance the pleasure of the person. The music is being classified in three different genres and based on established studies; we have chosen three different aromas to be released based on music genre. The intensity of the aroma released will be based on the average tempo of the music. With Ensemble Classification technique (ANN and k-Means), the classification rate of Music-Speech classifier for both Music and Speech was brought in range of 90%. An experimental study was conducted to analyse the effects of Aroma on Human Physical and Mental performance and his relaxation. The results attained were similar to those reported in literature.

Part B:

Linear and Non-Linear Global feature based classification of Emotional Speech

ABSTRACT

Speech emotion analysis refers to the use of various methods to analyse vocal behaviour as a marker of affect (e.g., emotions, moods, and stress), focusing on the nonverbal aspects of speech. The basic assumption is that there is a set of objectively measurable voice parameters that reflect the affective state a person is currently experiencing (or expressing for strategic purposes in social interaction). In this report, emotional speech recognition has been reviewed with two goals. The first is to present and analyse most frequent acoustic features used for emotional speech recognition and to assess how the emotion affects them. Typical features are the pitch, the formants, the vocal tract cross-section areas, the Mel-frequency Cepstral coefficients, the Teager energy operator-based features, the intensity of the speech signal, and the speech rate. The second goal is to review appropriate techniques in order to classify speech into emotional states. Artificial Neural Networks (ANN), Classification Trees and Support Vector Machines were used for classification of emotional speech.

CONTENTS

List of Figures

Part A: Music Genre/Tempo based Aroma Release System

Chapter3: Feature Set and Classification System

Chapter 4: Experiment to correlate Aromas with Relaxation, Physical and Mental Performance

Chapter 5: Algorithm Development and System Description

Chapter 6: Results

Part B:Linear and Non-Linear global Features based classification of Emotional Speech

Chapter 3: Speech Production Models and Feature Set

Chapter 4: Results

List of Tables

Part A: Music Genre/Tempo based Aroma Release System

Chapter 4: Experiment to correlate Aromas with Relaxation, Physical & Mental Performance

Chapter 6: Results (Matlab Results)

Embedded System Results

Part B:Linear and Non-Linear global Features based classification of Emotional Speech

Chapter 3: Speech Production Models and Feature Set

Chapter 4: Results

List of Flowcharts

Part A: Music Genre/Tempo based Aroma Release System

Chapter3: Feature Set and Classification System

Chapter 5: Algorithm Development and System Description

Acronyms and Abbreviations

AM	Amplitude Modulated
ANN	Artificial Neural Networks
ANS	Autonomous Nervous System
ASR	Automatic Speech Recognition
Auto	Autocorrelation
BB	Beagle-Bone Black
BPM	Beats per Minute
BW	Bandwidth
CF	Center Frequency
DC	Direct Current
DTFT	Discrete Time Fourier Transform
DWT	Discrete Wavelet Transform
EEG	Electroencephalography
FFT	Fast Fourier Transform
FM	Frequency Modulated
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
Hz	Hertz
kNN	k-Nearest Neighbour
LP	Linear Prediction
LPC	Linear Predictive Coding
LPCC	Linear Prediction Cepstral coefficients
MFB	Mel Filter Bank
MFCC	Mel-Frequency Cepstral Coefficients
OT	Other Than
PCB	Printed Circuit Board
pdf	Probability Distribution function
PDM	Power Distribution Module
RMS	Root Mean Square
SD	Standard Deviation

SFS	Sequential Forward Selection
SS	Single Stage
ST	Short Time
STFT	Short Time Fourier Transform
SVM	Support Vector Machine
TEO	Teagre Energy Operator
TS	Two Stage
TSST	Trier Social Stress Test
TV	Tele-Vision
Var	Variance
V/V	Volume by Volume
ZCR	Zero Cross Rate

Part A

Music Genre/Tempo based Aroma Release System

Chapter 1: Introduction

1.1: Music and Human Society

"Music is so naturally united with us that we cannot be free from it even if we so desired" (Boethius cited by Storr).

Music's interconnection with society can be seen throughout history. Every known culture on the earth has music. Music seems to be one of the basic actions of humans. Music affects the brain in many positive ways. It makes you smarter, happier and more productive at any age. It has been proven that music influences humans both in good and bad ways. These effects can be instant or long lasting. Music is thought to link all of the emotional, spiritual, and physical elements of the universe. Music can also be used to change a person's mood, and has been found to cause similar physical responses in many people simultaneously. Music also has the ability to strengthen or weaken emotions from a particular event such as a funeral. People perceive and respond to music in different ways. The level of musicianship of the performer and the listener as well as the manner in which a piece is performed affects the "experience" of music. An experienced and accomplished musician might hear and feel a piece of music in a totally different way than a non-musician or beginner. This is why two accounts of the same piece of music can contradict themselves [21], [22]. Rhythm is an important aspect of music, when looking at responses to music. There are two responses to rhythm. These responses are hard to separate because they are related, and one of these responses cannot exist without the other [22], [23], and [24]. These responses are

- The actual hearing of the rhythm and
- The physical response to the rhythm.

Rhythm organizes physical movements and is very much related to the human body. For example, the body contains rhythms in the heartbeat, while walking, during breathing, etc.

1.2: Aroma's Effect

Odours do affect people's mood, work performance and behaviour in a variety of ways. Odours don't work on us like a drug, instead we work on them through our experiences with them. That is, in order for an odour to elicit any sort of response in you, you have to first learn to associate it with some event. In olfaction, the process can be understood as follows: a novel odour is experienced in the context of an unconditioned stimulus, such as surgical procedure in a hospital, which elicits an unconditioned emotional response, such as anxiety [25].

A number of studies have shown that the odours people like make them feel good, whereas odours people dislike make them feel bad. These mood responses have also been reported physiologically. For example, skin conductance, heart-rate and eye-blink rates in response to various liked or disliked scents coincide with the mood the person is experiencing. Downstream from how odours influence our moods is the way that moods influence how we think (cognition) and how we act (behaviour). In terms of cognition, mood has been shown to influence creativity with the typical finding that people in a positive mood exhibit higher levels of creativity than individuals in a bad mood. Odours can also produce the same effects. When people were exposed to an odour they liked creative problem solving was better than it was when they were exposed to an unpleasant odour condition.

1.3: Music Genres

A music genre is a conventional category that identifies some pieces of music as belonging to a shared tradition or set of conventions. It is to be distinguished from musical form and musical style. Music can be divided into different genres in many different ways. The artistic nature of music means that these classifications are often subjective and controversial, and some genres may overlap. A music genre or subgenre may also be defined by the musical techniques, the style, the cultural context, and the content and spirit of the themes. Geographical origin is sometimes used to identify a music genre, though a single geographical category will often include a wide variety of subgenres. Among the criteria often used to classify musical genres are the tracheotomy of art, popular, and traditional music.

Art Music: The term art music refers primarily to classical traditions, including both contemporary and historical classical music forms [28]. Art music exists in many parts of the world. In Western practice, art music is considered primarily a written musical tradition, preserved in some form of music notation rather than being transmitted orally, by rote, or in recordings, as popular and traditional music usually are.

Popular Music: The term popular music refers to any musical style accessible to the general public and disseminated by the mass media [26]. Popular music, unlike art music, is:

- Conceived for mass distribution to large and often sociocultural heterogeneous groups of listeners,
- Stored and distributed in non-written form,
- Only possible in an industrial monetary economy where it becomes a commodity

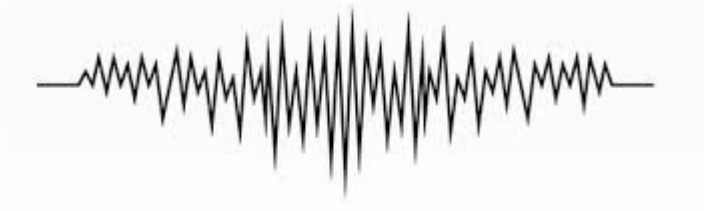
Traditional Music: Traditional music is a modern name for what has been called "folk music", excluding the expansion of the term folk music to include much non-traditional material. The two are both unified as traditional music due to:

- Oral transmission: The music is handed down and learned through singing, listening, and sometimes dancing;

- Cultural basis: The music derives from and is part of the traditions of a particular region or culture.

Chapter 2: Literature Review

2.1: Music–Speech classification



As Automatic Speech Recognition (ASR) systems are applied more and more in real world multimedia domains, the problem of speech-music classification has become important [1].

- To apply Automatic Speech Recognition (ASR) on music track we need to differentiate portions that contain speech and those which don't [2].
- Low bit-rate audio coding: [1] traditionally, separate codec designs are used to digitally encode speech and music signals. In many emerging multimedia applications such as the Internet, the sound stream carries both speech and music. Designing a universal coder to reproduce well both speech and music is the best approach.
- Third and a more common use could be to design a low cost music-speech classifier to shift the radio channels automatically to a channel where music is being played, rather than listen to chat and/or advertisement [1].

2.1.1: Differences between Music and Speech: In [4], the author discusses how music and speech can be differentiated

- **Tonality:** Music tends to be composed of a multiplicity of tones, each with a unique distribution of harmonics. This pattern is consistent regardless of the types of music or instruments. Speech exhibits an alternating sequence of tonal and noise-like segments.

- **Bandwidth:** Speech is usually limited in frequency to about 8 KHz whereas music can extend through the upper limits of ear's response at 20 KHz. In general, most of the signal power in music waveforms is concentrated at lower frequencies.

- **Excitation Patterns:** The excitation signals (pitch) for speech usually exist only over a span of three octaves while the fundamental music tones can span up to six octaves.

- **Tonal duration:** The duration of vowels in speech is very regular, following a syllabic rate. Music exhibits a wider variation in tone length, not being constrained by the process of articulation. Hence, tonal duration would likely be a good discriminator.

- **Energy Sequences:** A reasonable generalization is that speech follows a pattern of high energy conditions of voicing followed by low energy conditions which the envelope of music is less likely to exhibit.

2.1.2: Most commonly used features: Based on [1], [2], [3], [4], the most commonly used features are

- **Percentage of Low energy frames:** defines percentage of frames have RMS power less than 50% of mean of RMS power over one second window.
- **Spectral Roll-off point:** The 95^{th} percentile of power spectral distribution
- **Spectral Centroid:** The balancing point of spectral power.
- **Spectral Flux**
- **Zero cross rate:** The number of time domain zero crossing within a speech frame.

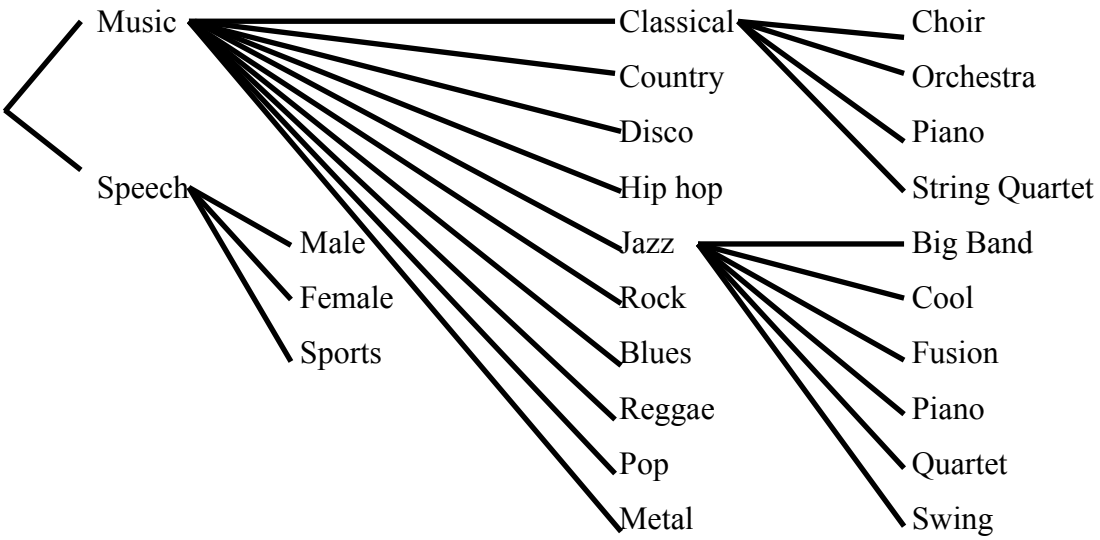
Gaussian Mixture models, Artificial Neural Networks and kNN were used to classify the sound segment to music or speech. The accuracy obtained varied from 80 to 98%. In [1], the accuracy attained was 94.6% based on frame by frame basis and 98.6% when used over long window size. In [2], using mean, variance, and time averages of spectral features, they attained an accuracy varying 70% to 90% for various different genres, using k-Means and Gaussian quadratic classifiers. In [3], the author achieved an accuracy of 90% for music and 94% for speech detection based on Gaussian classifiers. In [4], the system was tested on FM radio providing accuracy of 98%. Mean and variance of the features described above were used to generate a feature vector for a sound segment.

2.2: Music Genre classification: Music genres are labels created and used by humans for categorizing and describing the vast universe of music. This observation has led some researchers to suggest the definition of a new genre classification scheme purely for the purposes of music information retrieval.

2.2.1: Categories of features that could be used: In [5], the author discusses the categories of features explored by various authors for music genre classification.

- **Instrumentation:** e.g. whether modern instruments are present
- **Musical Texture:** e.g. Standard deviation of average melodic leap of different lines
- **Rhythm:** e.g. average time between attacks
- **Dynamics:** e.g. average note to note changes in loudness
- **Pitch Statistics:** e.g. fraction of notes in the bass register
- **Melody:** e.g. fraction of melodic intervals comprising a tritone
- **Chords:** e.g. prevalence of most common vertical interval

2.2.2: Audio classification hierarchy



2.2.3: Features Set and Results: Feature extraction is the process of computing a compact numerical representation that can be used to characterize a segment of audio. The design of descriptive features for a specific application is the main challenge in building pattern recognition systems. Once the features are extracted standard machine learning techniques which are independent of the specific application area can be used. In [7], the author described the overall standard feature set used for music genre classification.

A. **Timbral Texture Features:** The calculated features are based on the Short Time Fourier transform (STFT) and are calculated for every short-time frame of sound [7].

- **Spectral Centroid:** The spectral centroid is defined as the center of gravity of the magnitude spectrum of the STFT.

- **Spectral Roll off:** The spectral roll off is defined as the frequency below which 85% of the magnitude distribution is concentrated.

- **Spectral Flux:** The spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions.

- **Time Domain Zero cross rate:** Time domain zero crossings provide a measure of the noisiness of the signal.

- **Mel-Frequency Cepstral Coefficients:** After taking the log-amplitude of the magnitude spectrum, the FFT bins are grouped and smoothed according to the perceptually motivated Mel-frequency scaling. Finally, in order to de-correlate the resulting feature vectors a discrete cosine transform is performed.

- **Low Energy Feature:** It is defined as the percentage of analysis windows that have less RMS energy than the average RMS energy across the texture window.

B. **Rhythmic Content Features:** The regularity of the rhythm, the relation of the main beat to the sub-beats, and the relative strength of sub-beats to the main beat will be represented through feature vectors [7],[16],[18].

Gaussian Mixture models, Artificial Neural Networks, K-nearest neighbour and SVM were used to classify the music segments of length varying from 1.2 seconds to 30 seconds to various genres. The accuracy obtained varied from 45 to 70% for various genres. In [7], the author reported an accuracy of 61%, which is comparable to human classification. In [8], the author used MFCC to classify music genres and attained an accuracy of 70 to 805 on 1500 sound segments. In [9], the author used Support Vector Machines (SVM) to classify the music genres using a test set of 60,000 and attained a accuracy of 80% for binary classification. In [13] the author presents a classification hierarchy model for 29 genres, dividing them in groups, and subgroups achieving an accuracy of 70 to 80% for various stages.

2.3: Aroma and its Effect:

In [28], "odours and consumer behaviour in a restaurant", the author proposes hypothesis that lavender produces a soothing effect on a person. Increase in performance of difficult cognitive task, when diffusing a pleasant smell (floral) was observed in [33]. In [29] the author observed increased performance in math computation task. In [34], the author validated the increase in physical performance through peppermint odour. In [29], "Enhancing athletic performance through the administration of peppermint odour", they conducted an experiment on 40 persons, who took physical tests in presence and absence of peppermint odour. The peppermint odour resulted in increased running speed, hand grip, and no of push-ups. In, "The effects of peppermint on exercise performance" [30], the results of the experiment support the effectiveness of peppermint essential oil on the exercise performance, gas analysis, spirometer parameters, blood pressure, and respiratory rate in the young male students. Relaxation of bronchial smooth muscles, increase in the ventilation and brain oxygen concentration, and decrease in the blood lactate level were presented as the most plausible explanations.

In "The effects of lavender oil inhalation on emotional states, autonomic nervous system, and brain electrical activity", [31] the author analysed blood pressure, heart rate, skin temperature, and respiratory rate of the subjects. The results revealed that lavender oil caused significant decreases of blood pressure, heart rate, and skin temperature, which indicated a decrease of autonomic arousal. In terms of mood responses, the subjects in the lavender oil group categorized themselves as more active, fresher, relaxed than subjects just inhaling base oil. Compared with base oil, lavender oil increased the power of theta (4-8 Hz) and alpha (8-13 Hz) brain activities. The topographic map showed obviously more scattering power in alpha range waves particularly in bilateral temporal and central area.

2.4: Music Effects on Human Brain: In, "Sensation seeking and music preferences." [36], the author hypothesis that Thrill, Adventure and Experience seekers have a preference for Classical, Folk and Rock Music. Also, high sensation seekers have a high optimal level of stimulation and thus tolerate/like high intensity and/or complexity in music, and stimulation in general. In [37] "Listening to Mozart enhances spatial-temporal reasoning: towards a neurophysiological basis", the author performed a behavioural experiment which showed that listening to a Mozart piano sonata produced significant short-term enhancement of spatial-temporal reasoning in college students. They also hypothesised and proved that repetitive' music and a taped short story does not enhance reasoning. In [38] "Hits to the left, flops to the right: different emotions during listening to music are reflected in cortical lateralisation patterns.", the author investigates the neurobiological mechanisms accompanying emotional valence judgements during listening to complex auditory stimuli, cortical direct current (dc)-electroencephalography (EEG) activation patterns were recorded from 16 right-handed students. Students listened to 160 short sequences taken from the repertoires of jazz, rock-pop, classical music and environmental sounds (each n = 40). Emotional valences of the perceived stimuli were rated on a 5-step scale after each sequence. Brain activation patterns during listening revealed widespread bilateral fronto-temporal activation, but a highly significant lateralisation effect: positive emotional attributions were accompanied by an increase in left temporal activation, negative by amore bilateral pattern with preponderance of the right fronto-temporal cortex. Female participants demonstrated greater valence-related differences than males. No differences related to the four stimulus categories could be detected, suggesting that the actual auditory brain activation patterns were more determined by their affective emotional valence than by differences in acoustical "fine" structure. The results are consistent with a model of hemispheric specialisation concerning perceived positive or negative emotions proposed by Heilman.

In [39] "Does music enhance cognitive performance in healthy older adults? The Vivaldi effect", it was found that Classical music significantly increased working memory performance compared with the no-music condition. The arousal-and-mood hypothesis [43] claims that music enhances the level of arousal, and consequently attention processes benefit, and/or that it promotes positive mood. In particular, the theory holds that adding entertaining auditory backgrounds makes the learning task more interesting and thereby

increases the learner's overall level of arousal. This increase in arousal results in a greater level of attention, so that more material is processed by the learner, resulting in improved performance on retention tests. In [40], "Effects of Progressive Relaxation and Classical Music on Measurements of Attention, Relaxation, and Stress Responses", the four groups of subjects exhibited similar performance on behavioural measures of attention that suggested a reduction in physiological arousal following their relaxation or control condition, as well as a decreased heart rate. Progressive Relaxation, however, resulted in the greatest effects on behavioural and self-report measures of relaxation, suggesting that cognitive cues provided by stress management techniques contribute to relaxation.

In [41], "Impact of Music on Brain Function during Mental Task using Electroencephalography", the objective of this study was to analysis the effect of music (carnatic, hard rock and jazz) on brain activity during mental work load using electroencephalography. Spectral powers features were extracted at alpha, theta and beta brain rhythms. While listening to jazz music, the alpha and theta powers were significantly ($p < 0.05$) high for rest as compared to music with and without mental task in Cz. While listening to Carnatic music, the beta power was significantly ($p < 0.05$) high for with mental task as compared to rest and music without mental task at Cz and Fz location. This finding corroborates that attention based activities are enhanced while listening to jazz and Carnatic as compare to Hard rock during mental task. In [45], the author claims that the psychological and physiological health of individuals can be improved by music therapy. In [46], Neurological studies have establishes that music is a valuable tool for evaluating the brain system. In [42], "Music and Emotions in the Brain: Familiarity Matters", the author conducts a study to clarify the role of familiarity in the brain correlates of music appreciation by controlling, in the same study, for both familiarity and musical preferences. Brain activation data revealed that broad emotion-related limbic and Para limbic regions as well as the reward circuitry were significantly more active for familiar relative to unfamiliar music. Smaller regions in the cingulate cortex and frontal lobe, including the motor cortex and Broca's area, were found to be more active in response to liked music when compared to disliked one. He concludes that familiarity seems to be a crucial factor in making the listeners emotionally engaged with music, as revealed by fMRI data.

Chapter 3: FEATURE SET AND CLASSIFICATION SYSTEM

3.1: Dataset used: The data used must represent as much of the breadth of available input signals as possible. The MARSYS dataset has both male and female speakers, both "in the studio" and telephonic, with quiet conditions and with varying amounts of background noise in the speech class; and samples of jazz, pop, country, salsa, reggae, classical, various non-Western styles, various sorts of rock, and new age music, both with and without vocals in the music class.

3.1.1: Music-Speech Classification: The dataset used for music-speech classification was downloaded from free open source MARSYS, Music analysis, retrieval and synthesis for audio signals. (http://marsyas.info/downloads/datasets.html). The dataset contains 120 tracks, each 30 seconds long and each class of music and speech has 60 examples each. The data was sampled at rate of 22050Hz.

3.1.2: Music Genre Classification: The dataset used for music genre classification was downloaded from free open source MARSYS, Music analysis, retrieval and synthesis for audio signals. (http://marsyas.info/downloads/datasets.html). The dataset contains 1000 tracks, each 30 seconds long for ten different genres. The data was sampled at rate of 22050Hz.

3.2: Features Set

- **Spectral Centroid:** The spectral centroid is defined as the center of gravity of the magnitude spectrum of the STFT

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]} \qquad \textbf{Eq. 3.2.1}$$

where

$M_t[n]$ *is the magnitude of Fourier transform at frame* t *and frequency bin* n.

The centroid is a measure of spectral shape and higher centroid values correspond to "brighter" textures with more high frequencies.

- **Spectral Roll off:** The spectral roll-off is defined as the frequency R_t below which 85% of the magnitude distribution is concentrated

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \qquad \boldsymbol{Eq.3.2.2}$$

The roll-off is another measure of spectral shape.

- **Spectral Flux:** The spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \qquad \boldsymbol{Eq.3.2.3}$$

$where\ N_t[n]\ and\ N_{t-1}[n] are\ the\ normalized\ magnitude\ of\ fourier\ transform\ at\ the\ current\ frame\ t\ and\ the\ previous\ frame\ t-1, respectively.$

The spectral flux is a measure of the amount of local spectral change.

- **Time Domain Zero Cross Rate:**

$$Z_t[n] = \frac{1}{2}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \qquad \boldsymbol{Eq.3.2.4}$$

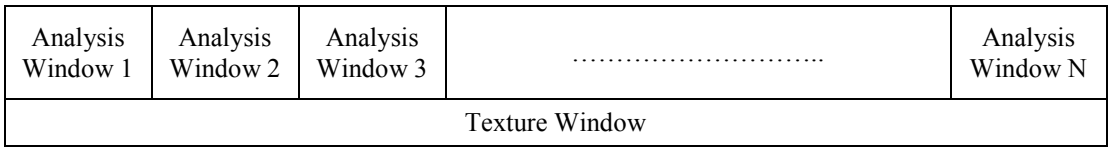
$where\ the\ sign\ function\ is\ 1\ for\ positive\ arguements\ \ and\ 0\ for\ negative\ arguements\ and\ x[n] is\ \ the\ time\ domian\ signal\ for\ frame\ t.$

Time domain zero crossings provide a measure of the noisiness of the signal.

- **Mel-Frequency Cepstral Coefficients:** Mel-frequency Cepstral coefficients (MFCC) are perceptually motivated features that are also based on the STFT. After taking the log-amplitude of the magnitude spectrum, the FFT bins are grouped and smoothed according to the perceptually motivated Mel-frequency scaling. Finally, in order to de-correlate the resulting feature vectors a discrete cosine transform is performed.

- **Analysis and Texture Window:** In short-time audio analysis, the signal is broken into small, possibly overlapping, segments in time and each segment is processed separately. These segments are called analysis windows and have to be small enough so that the frequency characteristics of the magnitude spectrum are

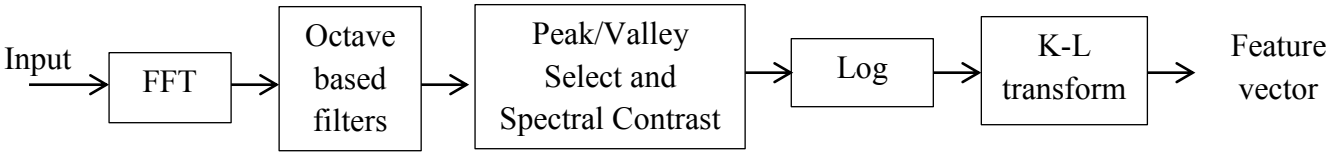
relatively stable (i.e., assume that the signal for that short amount of time is stationary). However, the sensation of a sound "texture" arises as the result of multiple short-time spectrums with different characteristics following some pattern in time. For example, speech contains vowel and consonant sections which have very different spectral characteristics. Therefore, in order to capture the long term nature of sound "texture," the actual features computed in our system are the running means and variances of the extracted features described in the previous section over a number of analysis windows.



- **Low Energy Feature:** Low energy is the only feature that is based on the texture window rather than the analysis window. It is defined as the percentage of analysis windows that have less RMS energy than the average RMS energy across the texture window.

To summarize, the feature vector for describing Timbral texture consists of the following features: means and variances of spectral centroid, roll-off, flux, zero crossings over the texture window, low energy frames, and means and variances of the first five MFCC coefficients over the texture window.

- **Spectral Contrast Feature:** Octave-based Spectral Contrast considers the spectral peak, spectral valley and their difference in each sub-band. For most music, the strong spectral peaks roughly correspond with harmonic components; while non-harmonic components, or noises, often appear at spectral valleys. Thus, Spectral Contrast feature could roughly reflect the relative distribution of the harmonic and non-harmonic components in the spectrum.



Octave-based Spectral Contrast uses octave-scale filter, while MFCC uses Mel-scale filters. Although Mel-scale is suitable for general auditory model, octave-scale filter is more suitable for music processing.

3.3: Beat Detection

In music and music theory, the beat is the basic unit of time, the pulse (regularly repeating event), of the *mensural level* (or *beat level*). The beat is often defined as the rhythm listeners would tap their toes to when listening to a piece of music, or the numbers a musician counts while performing. In popular use, *beat* can refer to a variety of related concepts including: tempo, meter, specific rhythms, and groove. Rhythm in music is characterized by a repeating sequence of stressed and unstressed beats (often called "strong" and "weak") and divided into bars organized by time signature and tempo indications. Metric levels faster than the beat level are division levels, and slower levels are multiple levels.

The lowest level of short vertical lines mark the location of each sixteenth-note sub-beat within one measure of 4/4 time. There are sixteen of these locations in the bar. Moving up the diagram, we see that a quarter-note (a beat) is four times the length of a sixteenth-note. Four quarter notes make up the length of the whole measure, just as 16 sixteenth-notes do.

Fig 3.3.1: *Beat, Sub-beat Description diagram*

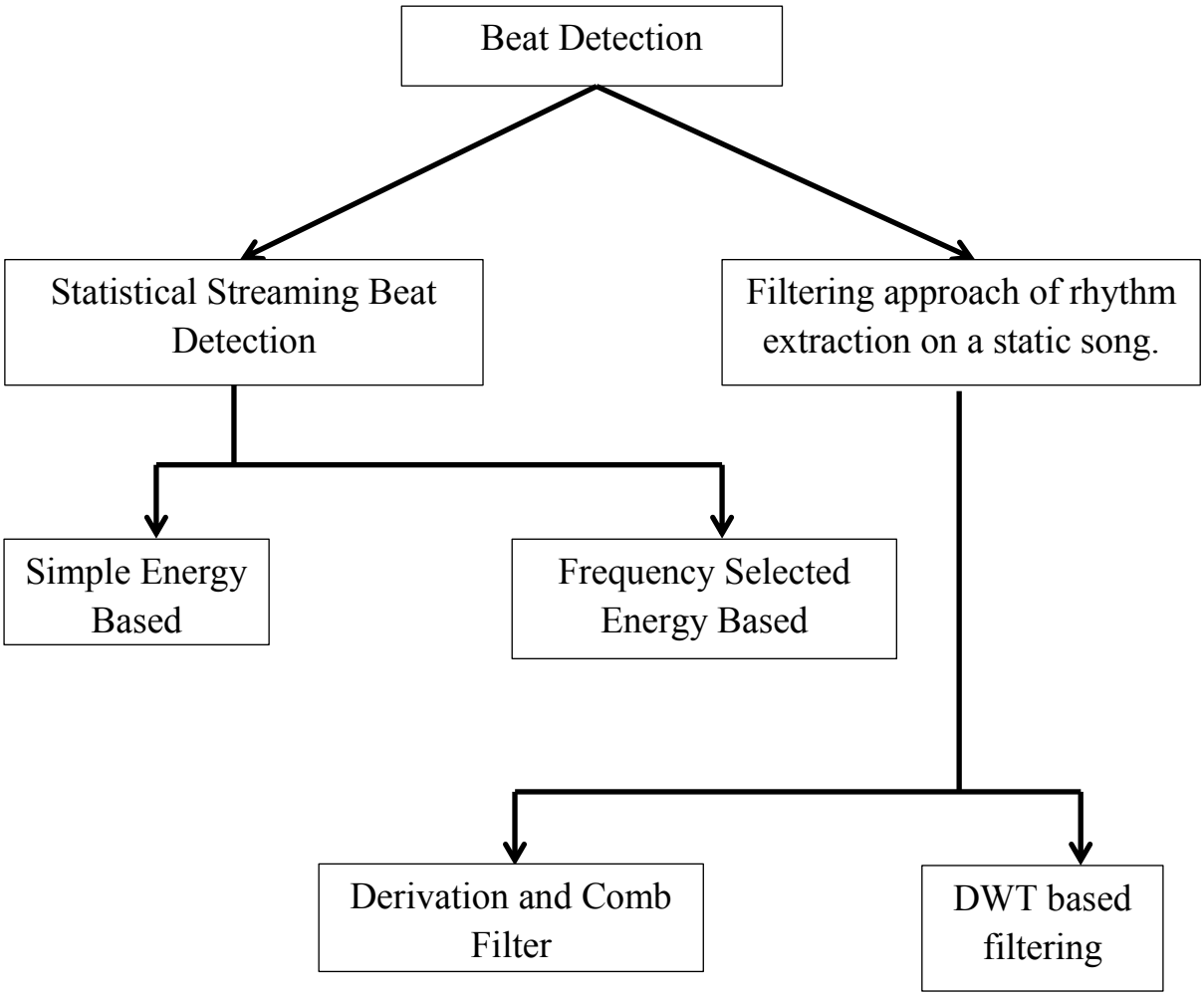


Fig 3.3.2: *Beat Detection Algorithm*

3.3.1: Simple Sound Energy Based

The human listening system determines the rhythm of music by detecting a pseudo – periodical succession of beats. The signal which is intercepted by the ear contains certain energy; this energy is converted into an electrical signal which the brain interprets. Obviously, the more energy the sound transports, the louder the sound will seem. But a sound will be heard as a beat only if its energy is largely superior to the sound's energy history that is to say if the brain detects a brutal variation in sound energy. Therefore if the ear intercepts a monotonous sound with sometimes big energy peaks it will detect beats, however, if you play a continuous loud sound you will not perceive any beats. Thus, the

beats are big variations of sound energy. This first analysis brings us to our simplest model: *Sound energy peaks*.

Algorithm:

Step 1: Break the each second of the song in 43 windows of equal length. (for fs=44100, each window could have 1032 samples).

Step 2: Compute the energy of the current window

$$E_i = \sum_{k=0}^{1032} (x[k])^2 \qquad \textbf{Eq. 3.3.1}$$

Step 3: If $E_i > C \times mean(\ last\ 43\ windows\ energy)$, *then we have a beat.*

Repeat this for all the windows in the song. And also keep updating the mean of energy of last 43 windows.

Calculation of C: *Calculate the variance of energy of last 43 windows (V). The value of C is given by*

$$C = (-0.0000015 \times V) + 1.5142857 \qquad \textbf{Eq. 3.3.2}$$

This beat detection is very accurate and sounds right with techno and rap, the beats are very precise and the music contains very little noise. The beat detection on punk, rock and hard rock, is sometimes quite approximate. This algorithm fails to get the rhythm of the song. Indeed the algorithm detects energy peaks. Sometimes you can hear a drum beat which is sank among other noises and which goes through the algorithm without being detected as a beat.

3.3.2: Frequency Selected Energy Based

This algorithm has the ability to determine on which frequency sub-band we have a beat and if it is powerful enough to take it into account. Basically it tries to detect big sound energy variations in particular frequency sub-bands and separate beats regarding their frequency sub-band. If required, more importance could be given to low frequency beats or to high frequency beats.

Algorithm:

Step 1: Break the each second of the song in 43 windows of equal length. (for fs=44100, each window could have 1032 samples).

Step 2: Compute the fft of each window

Step 3: Break each window into 32 sub-bands where each sub-band contains 32 samples of fft of the window.

Step 4: Compute the sum of squares of the amplitude of the complex numbers in the sub-band E_{si}, where s denotes the sub-band and I denotes the window number.

$$E_{si} > C \times mean\ of\ last\ 43\ energies\ of\ this\ particular\ subband,$$
$$then\ we\ have\ a\ beat\ in\ this\ subband.$$

Here value of C = 15.6

To improve above algorithm, increase the number of sub-bands from 32 to 64. This takes more computing time but it also gives more precision in our beat detection. The second way to develop the accuracy of the algorithm uses the defaults of human ears. Human hearing system is not perfect; in fact its transfer function is more like a low pass filter. We hear more easily and more clearly low pitched noises than high pitch noises. This is why it is preferable to make a logarithmic repartition of the sub-bands. That is to say that sub-band 0 will contain only say 2 frequencies whereas the last sub-band, will contain say 20.

3.3.3: Beat Detection Discrete Wavelet Transform (DWT): Unlike the STFT that provides uniform time resolution for all frequencies the DWT provides high time resolution and low frequency resolution for high frequencies and high frequency resolution and low time resolution for low frequencies. In that respect it is similar to the human ear which exhibits similar time-frequency resolution characteristics. The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently. The DWT is defined as:

$$W(i,j) = \sum_j \sum_k x(k) 2^{-\frac{j}{2}} \varphi\left(2^{-j} n - k\right) \qquad \boldsymbol{Eq.3.3.3}$$

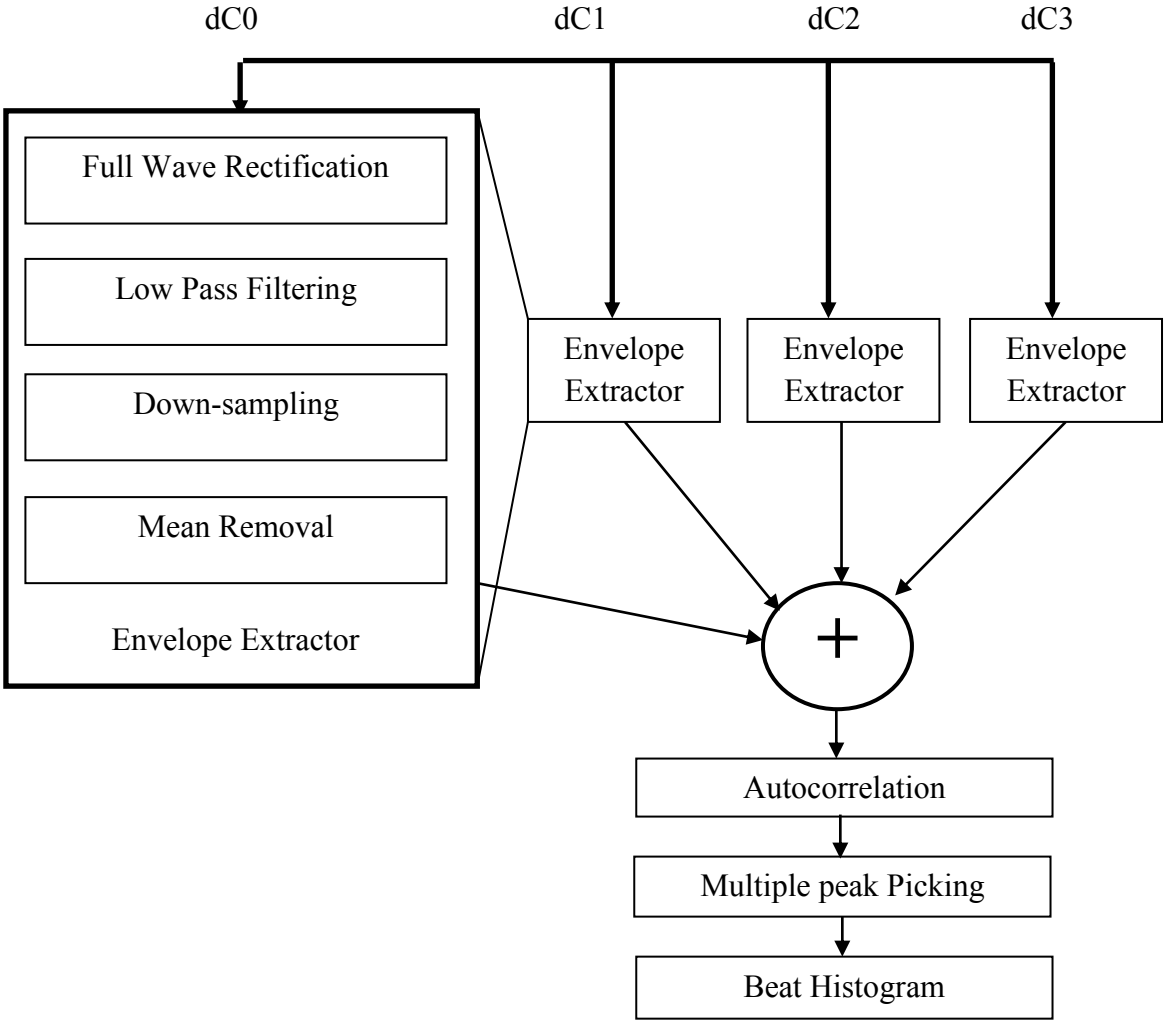
where $\varphi(t)$ is a time function with finite energy and fast decay called mother wavelet. As a multi-rate filter-bank the DWT can be viewed as a constant Q filter-bank with octave spacing between the centres of the filters. Each sub-band contains half the samples of the neighbouring higher frequency sub-band. In pyramidal algorithm the signal

is analysed at different frequency bands with different resolution by decomposing the signal into a coarse approximation and detail information. The coarse approximation is achieved by successive high-pass and low-pass filtering of the time domain signal and is defined by the following equations:

$$y_{high}[k] = \sum_{n} x[n]g[2n-k] \qquad \boldsymbol{Eq.3.3.4}$$

$$y_{low}[k] = \sum_{n} x[n]h[2k-n] \qquad \boldsymbol{Eq.3.3.5}$$

where $y_{high}[k], y_{low}[k]$ are the outputs of the highpass(g) and lowpass(h) filters, respectively after subsampling by 2.



Flow-Chart 3.3.1: *DWT based Beat Detection Flow Diagram*

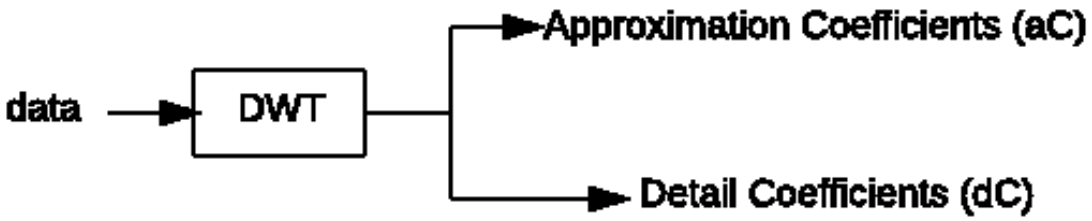


Fig 3.3.3: *DWT*

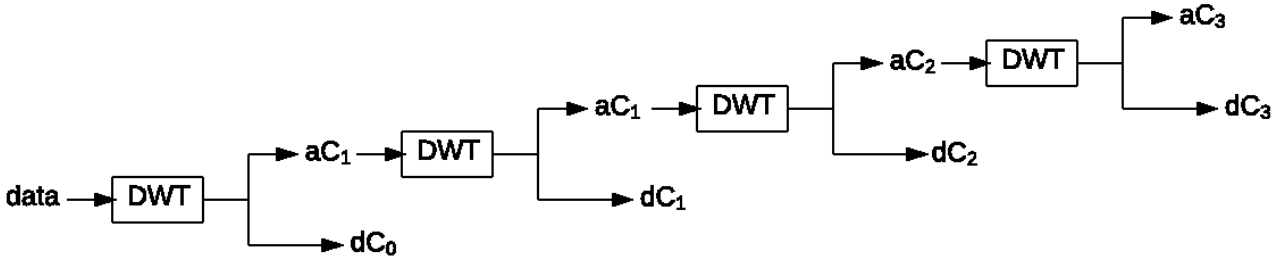


Fig 3.3.4: *Cascade DWT for beat detection*

3.3.1: Equations

- **Full Wave Rectification:**

$$y[n] = abs(x[n]) \qquad \boldsymbol{Eq.3.3.6}$$

- **Down-sampling**

$$y[n] = x[2n] \qquad \boldsymbol{Eq.3.3.7}$$

- **Normalization**

$$y[n] = x[n] - E(x[n]) \qquad \boldsymbol{Eq.3.3.8}$$

- **Autocorrelation**

$$y[k] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]x[n+k] \qquad \boldsymbol{Eq.3.3.9}$$

3.4: Audio Thumb-Nailing Using Chroma Based Representation [56]

For Meditation Assist, the system needs to detect the background noise and match to closest possible Natural sound, which is pre-stored in the system. The problem can reasonably be reduced to the problem of isolating repeated musical structures in an audio waveform. This algorithm makes use of a pattern recognition framework for audio streams, in which the signal is segmented into frames and each frame is described by a set of features.

CYCLIC REPRESENTATION OF FREQUENCY: In the early 1960s, Shepard reported that two dimensions rather than one are necessary to represent the perceptual structure of pitch [55]. He determined that the human auditory system's perception of pitch was better represented as a helix than as a one-dimensional line, and coined the terms tone height and Chroma to characterize the vertical and angular dimensions, respectively. Fig.3.4.1 shows an illustration of this helix with its two dimensions. In this representation, as the pitch of a musical note increases, say from C1 to C2, its locus moves along the helix, rotating chromatically through all of the pitch classes before it returns to the initial pitch class (C) one cycle above the starting point. According to Shepard's results, the perceived pitch, of a signal can be factored into values of Chroma, and tone height as

$$p = 2^{h+c}$$

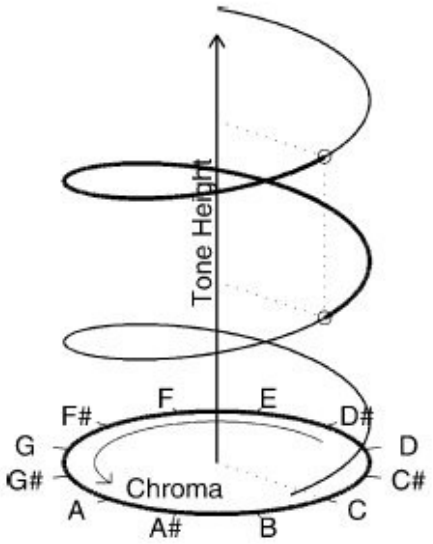


Fig 3.4.1: *Illustration of Shepard's helix of pitch perception. The vertical dimension is tone height, while the angular dimension is Chroma.*

For this decomposition to be unique it is sufficient for and linear changes in result in logarithmic changes in the fundamental frequency associated with the pitch. By dividing the interval between 0 and 1 into 12 equal parts, the 12 pitches of the equal-tempered chromatic scale can be obtained. The implication of Shepard's representation is that the distance between two pitches depends on both and, rather than on alone.

Algorithm:

1. Break the Audio signal in window size of 0.25 to 0.6 seconds.
2. For t^{th} frame, calculate the logarithmic magnitude of the DFT $\{F_t[n]\}$. The length of DFT is equal to the first power of 2 greater than or equal to length of the frame (or the longest frame if frame size varies, whichever is larger).
3. The elements of the Chroma feature vector for the t^{th} frame v_t are calculated using the equation

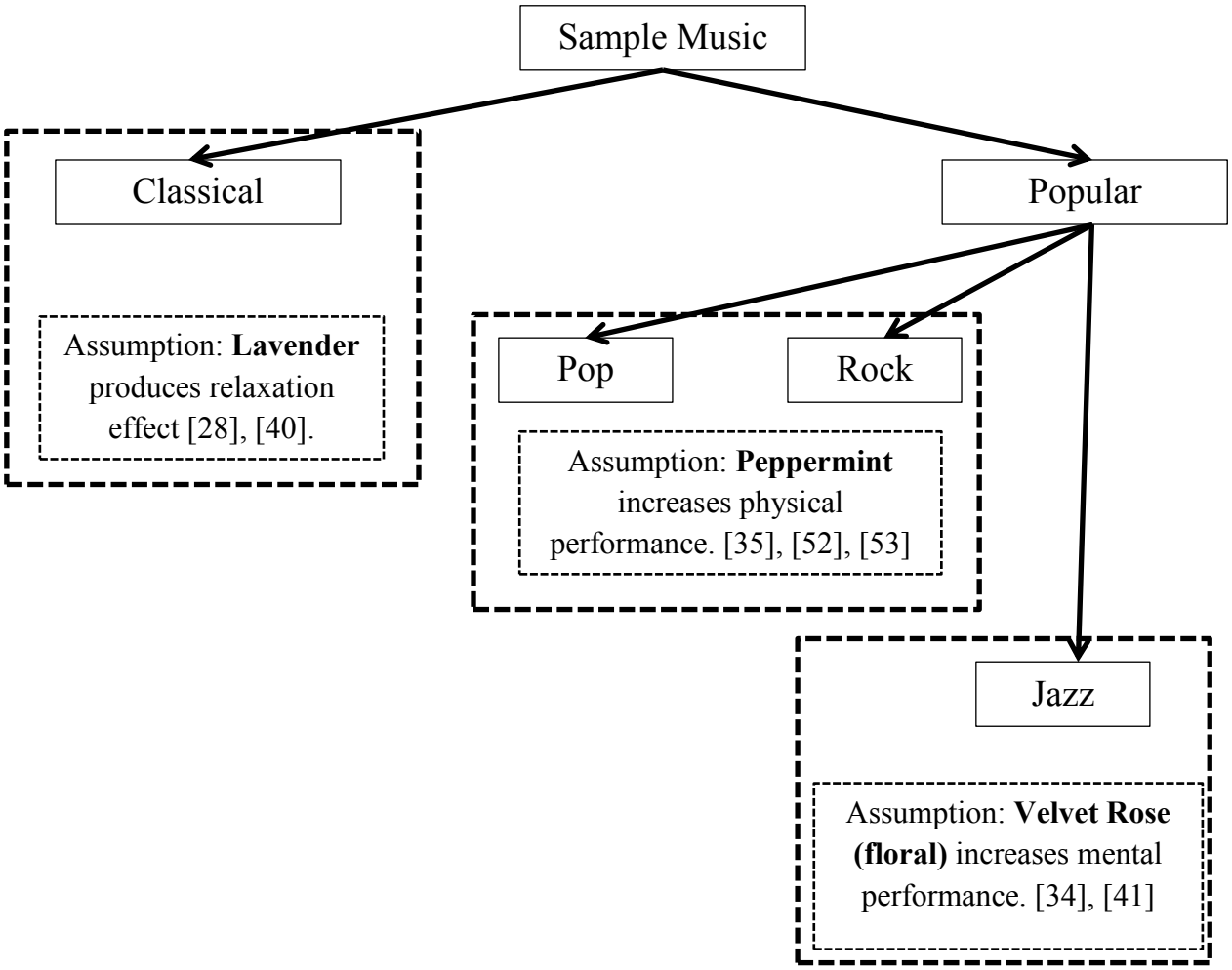
$$v_{t,k} = \sum_{n \in S_k} \frac{F_t[n]}{N_k}, \quad k \in \{0,1,2, \dots 11\} \qquad \boldsymbol{Eq.3.4.1}$$

   where each $S_k \in Z$, defines a subset of the all discrete frequency space for each pitch class and N_k is the number of elements in S_k .
4. The 12 sets S_k are generated by associating each DFT bin with one of the 12 pitch classes. Each bin's associated frequency *(f)* is calculated and then its Chroma is calculated using

$$c = log_2(f) - \lfloor log_2(f) \rfloor \qquad \boldsymbol{Eq.3.4.2}$$

5. Each bin 'i' is associated with Chroma value centred at Chroma values of $\frac{i}{12}$.
6. Lower bound of the spectrum is set at 20Hz to correspond to limit of Human Hearing. The upper bound is set at 2000Hz, because at this range at this frequency the critical bands of human auditory becomes broad enough to possibly admit multiple partial frequencies of a harmonic series.
7. For comparing two audio samples, calculate the Chroma Vector Matrix. Take the square difference sum over all its elements. The one with the least sum is closest to the audio signal.

Chapter 4: Experimentation to correlate Aromas' with Relaxation, Mental and Physical Performance



The purpose of the experiment conducted was to correlate Aromas with their effect to Human Mental and Physical Performance and his/her relaxation. The experiment was conducted in IIT Kharagpur, in RTES lab (Department of Electrical Engineering, IIT Kharagpur) and Language Lab (JCB Complex, IIT Kharagpur). The procedure of the experiment was discussed with Dr. Rajlakshmi Guha, Counsellor at Indian Institute of Technology, Kharagpur. Each participant was clearly explained the experiment procedure and was asked to sign an informed consent form (Detailed in Appendix A). Each participant was free to walk out of the experiment at any stage he/she feels uncomfortable. Overall twelve test subjects participated willingly in three different experiments.

4.1: Relaxation Experiment: In [28], the author proposes the hypothesis that; Lavender helps in relaxation and in [40], the author analysed the effect of classical Music on progressive relaxation. In his study, all four groups exhibited similar performance on behavioural measures of attention that suggested a reduction in physiological arousal following their relaxation or control condition, as well as a decreased heart rate. Progressive Relaxation, however, resulted in the greatest effects on behavioural and self-report measures of relaxation, suggesting that cognitive cues provided by stress management techniques contribute to relaxation. To correlate the effect of Lavender with relaxation, we conducted a Psychological Stress Induction test.

4.1.1: Psychological Stress Induction [48]: Stress will be induced in test subjects using Trier Social Stress Test (TSST). "The first 5-minute component is the anticipatory stress phase, during which the judges will ask the participant to prepare a 5-minute presentation. Also, the judges will maintain neutral expressions throughout the test. The participant is allowed to use paper and pen to organize their presentation, but this paper is then unexpectedly taken away from them when it is time to begin the presentation. During the 5-minute presentation component, the judges observe the participant without comment. If the participant does not use the entire 5 minutes, judge will ask him or her to continue. This goes on until the entire 5 minutes have been used. The presentation is immediately followed by the mental arithmetic component, during which the participant is asked to count backwards from 1,022 in steps of 13. If a mistake is made, then they must start again from the beginning. This component lasts for 5 minutes and is followed by a recovery period.

4.1.2: Test Subjects: Six selected participants all male aged between 18 and 24 years (mean age 20.5 $\pm$ 1.67 years) with normal body mass indices (mean 24.89 + 3.59) were enrolled in the present study. A summary of the demographic data of the participants is presented in Table 4.1.1. None of the subjects had abnormalities affecting smell, cardiovascular diseases, or a history of smoking or drug addiction. Twelve hours prior to testing subjects were asked to wash their hair without any spray. They were also asked not to use antiperspirants, perfumes and refrain from consuming alcohol, cigarettes, or

caffeinated drinks. They were requested to try to sleep well before the day of the experiment to avoid feeling fatigued or drowsy. Subjects were given a full explanation of the research and a written informed consent of all aspects of the present study, and were free to withdraw at any time.

Parameter	N	Minimum	Maximum	Mean	SD
Age (yrs)	6	18	23	20.5	1.64
Height (cms)	6	165	186	179.33	7.37
Weight (kgs)	6	68	102	80	12.18
Body Mass Index	6	21.38	31.48	24.89	2.59

Table 4.1.1: *Demographic data for the volunteers*

Fig 4.1.1: *Trier Social Stress Test Set-up*

4.1.3: Procedure: All experiments were conducted in a quiet room with ambient temperature (23 ± 1°C). The experiments were performed between 9.30 and 12.00 a.m. to minimize circadian variation. All participants attended to this research for two times, firstly, to measure the autonomic nervous system (Heart Rate and Blood Pressure only)

and EEG recording in absence of Lavender Aroma and secondly, in presence of Lavender Aroma while conducting Trier Social Stress Test. Before ANS measurement beginning, the researcher clearly informed the procedure, and then participants signed an Informed Consent Form describing the present study and their rights. Heart rate was recorded at one-minute intervals and Systolic and diastolic blood pressure was recorded every two minutes. The test consisted of three sessions. The first session was done to be used as reference. The second and third session took 20 minutes each. In the third, the Lavender air-wick was used in room.

4.1.4: Results: The mean and Standard Deviation (SD) values of autonomic parameters in the experiment are presented in Table 4.1.2. The data were compared on various autonomic parameters during absence of Lavender Aroma. In presence of Lavender Aroma subjects had significantly reduced Heart rate (all six had reduced Heart rate), Systolic Pressure reduced for all except one, while diastolic pressure reduced for three and almost equal for two, while one showed an increased diastolic pressure. Though the test subject size is quite small, because of time and monetary constraints, the results attained were in line with those predicted by others.

	Subject ID	HR Mean	HR SD	Systolic Mean	Systolic Std. Dev.	Diastolic Mean	Diastolic Std. Dev.
Without Aroma	sb 1	88.75	3.44	141.00	8.32	98.67	3.12
	sb 4	109.06	11.08	141.00	12.72	76.00	8.49
	sb 5	88.44	4.08	130.89	5.14	84.33	5.72
	sb 6	91.47	4.53	137.14	11.84	84.00	4.58
	sb 7	88.56	3.56	155.89	9.56	96.44	5.05
	sb 8	89.88	2.58	138.13	5.52	90.75	5.31
With Aroma	sb 1	84.88	3.03	138.00	8.92	97.38	3.78
	sb 4	102.40	5.82	130.17	5.35	77.17	2.32
	sb 5	87.94	3.17	125.83	4.93	87.67	3.78
	sb 6	84.80	3.23	127.13	5.95	77.13	5.36
	sb 7	84.13	2.23	144.44	6.77	90.89	3.62
	sb 8	80.94	2.74	138.44	4.16	85.33	9.25

Table 4.1.2: *Mean and Standard Deviation of ANS during Trier Social Stress Test*

EEG Data: An EEG recording for a time of 20 minutes was also recorded along with other parameters (as mentioned above), 15 minutes of the TSST and last 5 minutes to observe the relaxation pattern.

The set of 24 electrodes with 1 additional ground which was placed according to the international 10-20 system at FP1, FP2, PG1, PG2, FZ, F3, F4, F7, F8, T3, T4, T5, T6, C3, C4, CZ, PZ, P3, P4, PZ, O1, O2, A1 and A2. Electro-Caps are made of an elastic spandex type fabric with recessed, Silver/Silver-Chloride (Ag/AgCl) electrodes attached to the fabric. The online filter was set to a band pass with low pass is equal DC and high pass is equal 70 Hz. A/D rate was 250 Hz. Notch filter was open at 50 Hz. The relative power spectrums of the respective frequency bands derived by Fast Fourier Transformation (FFT) were expressed as follows:

Frequency bands	Bandwidth
Delta	(0-3.99 Hz)
Theta	(4-7.99 Hz)
Alpha1	(8-9.99 Hz)
Alpha2	(10-12.99 Hz)
Beta	(13-30 Hz).

Table 4.1.3: *Frequency bands for EEG*

EEG data Analysis: The EEG power was calculated for each frequency band among resting, sweet almond oil, and lavender oil inhalation. The studied areas were divided into the left anterior (Fp1, F3, F7), right anterior (Fp2, F4, F8), right posterior (P4, T6, O2), left posterior (P3, T5, O1), and middle (Fcz, Cz, Cpz) shown each band power with theta, alpha1, alpha2, Beta (Table 4.1.4) and shown in Fig 4.1.2. There were noticeable changes of band power in theta and alpha waves that significantly increased during the lavender inhalation in all brains areas.

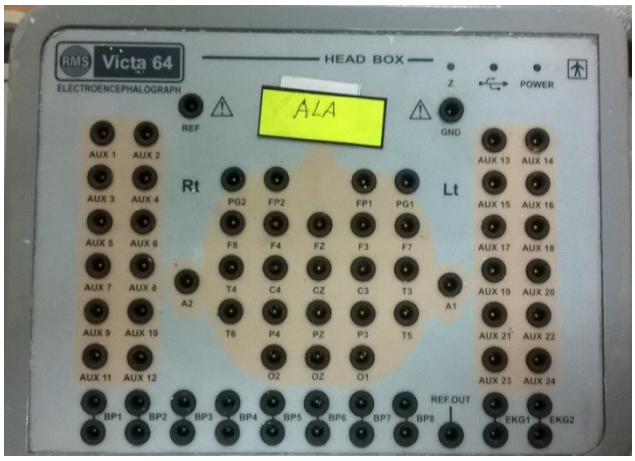


Fig 4.1.2: *EEG electrode position*

		Neutral	With Aroma	Without Aroma
Left Anterior	delta	0.379438	0.667619	0.764100
	theta	0.001197	0.021944	0.013813
	alpha1	0.391547	0.007466	0.511042
	alpha2	0.080185	0.004118	0.002925
	beta	0.478609	0.000988	0.605757
Right Anterior	delta	0.007571	0.874661	0.142056
	theta	0.000948	0.013948	0.012238
	alpha1	0.000217	0.367625	0.006199
	alpha2	0.000113	0.340488	0.004585
	beta	0.000038	0.374640	0.002504
Right Posterior	delta	0.149615	1.299807	0.143690
	theta	0.002182	0.101959	0.010877
	alpha1	0.128602	1.025883	0.040161
	alpha2	0.080485	0.073712	0.003567
	beta	0.305379	0.131056	0.608198
Left Posterior	delta	0.009813	0.968930	0.090000
	theta	0.001964	0.350111	0.010949
	alpha1	0.000917	0.728729	0.005673
	alpha2	0.000683	0.412888	0.003986
	beta	0.000288	0.269491	0.001139
Middle	delta	0.156519	1.375722	0.430064
	theta	0.002266	0.067243	0.016767
	alpha1	1.706811	0.876934	0.381357
	alpha2	1.449526	0.225789	0.178317
	beta	0.541611	0.867687	1.044538

Table 4.1.4: *Band power for theta, alpha1, alpha2, beta and delta frequency bands*

EEG Results: During inhalation with lavender, the power of theta (4-8 Hz) and alpha (8-13 Hz) activities are significantly increased in all brain regions. This result is consistent with the study of Diego [59] that found after lavender inhalation that frontal alpha power was significantly increased. The EEG evidence of relaxation can be seen in various practices such as meditation. Meditation is a way of balancing the body and the mind as well as controlling the mind to experience feelings of peace and relaxation.

	Day 1						Day 2 with aroma					
	sb1	sb4	sb5	sb6	sb7	sb8	sb1	sb4	sb5	sb6	sb7	sb8
Happy	xx	v	v	v	vv	v	v	vv	v	vv	v	vv
Sad	x	v	x	x	x	xx	xx	x	xx	x	xx	xx
Tired	v	v	v	v	x	v	xx	x	xx	x	xx	x
lively	x	v	vv	v	x	vv	v	vv	v	vv	v	v
caring	v	vv	vv	v	x	v	v	vv	vv	vv	x	v
content	v	vv	vv	v	vv	vv	vv	vv	v	vv	vv	vv
drowsy	x	xx	v	xx	x	xx	v	x	xx	xx	xx	xx
nervous	v	v	v	x	x	v	v	x	xx	x	x	xx
calm	x	v	v	v	vv	v	vv	v	v	v	vv	vv
Loving	x	v	vv	x	xx	vv	x	vv	v	x	vv	v
Fed up	x	xx	x	v	xx	xx	x	xx	xx	xx	xx	x
active	v	v	vv	v	vv	vv	vv	vv	vv	v	vv	vv

Table 4.1.5: Measuring Mood after Trier Social Stress Experiment

Positive Tired Mood Analysis [60]:

- Convert the Meddis response scale for Active, Caring, Lively, and Loving, to numbers this way: xx=1, x=2, v=3, vv=4. Add up the responses.
- Next, reverse score the responses for: Drowsy and Tired: xx = 4, x=3, v=2, vv=1. Add up the responses.
- Finally Add-up regular and reverse-regular added scores.

Positive Relaxed Mood Analysis [60]:

- Convert the Meddis response scale for Fed-up, Sad, and Nervous, to numbers this way: xx=4, x=3, v=2, vv=1. Add up the responses.
- Next, reverse score the responses for Calm: xx = 1, x=2, v=3, vv=4. Add up the responses.
- Finally Add-up regular and reverse-regular added scores.

Pleasant – Unpleasant Mood Analysis [60]:

- Convert the Meddis response scale for Drowsy, Fed up, Nervous, Sad, and Tired, to numbers this way: xx=4, x=3, v=2, vv=1. Add up the responses.
- Next, reverse score the responses for Active, Calm, Caring, Content, Happy, Lively, and Loving: xx = 1, x=2, v=3, vv=4. Add up the responses.
- Finally Add-up regular and reverse-regular added scores.

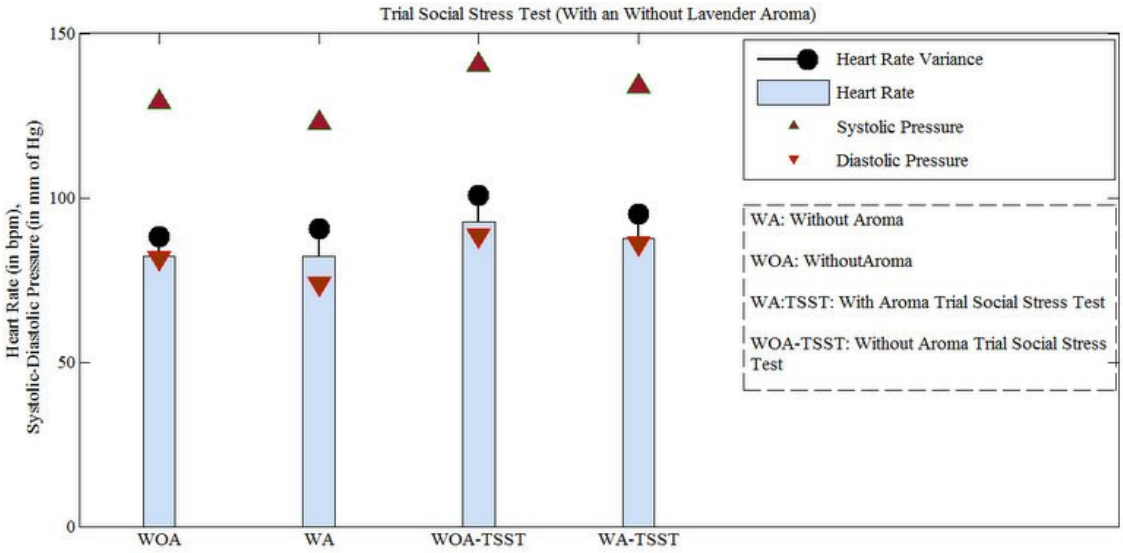


Fig 4.1.3: *Results for Stress Experiment*

4.1.5: Conclusions: Lavender produces a relaxing effect. Studies found that lavender increased drowsiness [54] and induced sleep. Therefore lavender seemed to relax people which in return led them to stay longer in the area where this smell was diffused. Naturally, this study has some limitations. The sample of test subjects tested was relatively small (n = 8). Therefore at this stage it would be impossible to generalize the results. The experiment needs to be conducted with a larger sample size.

- All six test subjects showed a reduced Heart Rate and less variance in each consecutive minute measurement.
- The Systolic Blood pressure fell 4.75% on an average, whereas Diastolic Blood Pressure showed a non-uniform relation.

Mood Analysis [60]: The subjects experienced more positive tired mood in case of presence of Lavender, compared to complete absence of Aroma. Similarly, the subjects experienced more relaxed mood in presence of Lavender. Also, the subjects in presence of Aroma (Lavender) scored more in case of Pleasant-Unpleasant Mood Analysis.

	With Lavender	Without Lavender
Pleasant-Unpleasant Scale	0.8715	0.7500
Positive Tired Mood	0.8541	0.7430
Positive Relaxed Mood	0.8750	0.7291

Table 4.1.6: *Mood Analysis*

4.2: Mental Performance Experiment: In [41], "Impact of Music on Brain Function during Mental Task using Electroencephalography", the author conducted an experiment to analyse the effect of Jazz, Carnatic and Rock music on Mental Task performance. He observed that while listening to jazz music, the alpha and theta powers were significantly ($p < 0.05$) high for rest as compared to music with and without mental task in Cz. While listening to Carnatic music, the beta power was significantly ($p < 0.05$) high for with mental task as compared to rest and music without mental task at Cz and Fz location. This finding corroborates that attention based activities are enhanced while listening to jazz and Carnatic as compare to Hard rock during mental task. In this experiment we try to validate the effect of floral (velvet Rose) aroma on enhancement of attention seeking mental activities.

4.2.1: Stroop Test: The Double Trouble task is based on phenomenon in the cognitive psychology literature known as the Stroop effect (Stroop, 1935). This effect refers to the increased difficulty one has in naming the print colour of a word, when the text of that word refers to an 'incongruent' colour. For example, people are slower to name the colour of red ink when the word that is written in red ink is the word 'green'. This difficulty in colour naming vanishes when the semantic meaning of the word is the same as the text colour (e.g. the word ' red ' written in red ink) or is a nonsense syllable (e.g. ' kyshqw ' written in red ink) and is diminished for semantically unrelated words (e.g. the word ' window ' written in red ink) [51]. This effect is thought to be the result of interference caused by automatic word recognition; it seems that we access the meaning of these words without consciously trying to do so. To perform this task successfully you must selectively focus your attention in order to inhibit the automatic access of distracting word information.

(Test Source: http://www.cambridgebrainsciences.com/browse/Reasoning/test/double-stroop-body)

4.2.2: Test Subject: Eight selected participants all male aged between 18 and 24 years (mean age 20.5 ± 1.67 years) with normal body mass indices (mean 24.89 + 3.59) were

enrolled in the present study. A summary of the demographic data of the participants is presented in Table 4.2.1. None of the subjects had abnormalities affecting smell, cardiovascular diseases, or a history of smoking or drug addiction. Twelve hours prior to testing subjects were asked to wash their hair without any spray. They were also asked not to use antiperspirants, perfumes and refrain from consuming alcohol, cigarettes, or caffeinated drinks. They were requested to try to sleep well before the day of the experiment to avoid feeling fatigued or drowsy. Subjects were given a full explanation of the research and a written informed consent of all aspects of the present study, and were free to withdraw at any time.

Parameter	N	Minimum	Maximum	Mean	SD
Age (yrs)	8	19	23	20.88	1.25
Height (cm)	8	150	183	170.13	12.78
Weight (kgs)	8	52	102	70.38	18.02
Body Mass Index	8	19.33	31.48	23.98	3.59

Table 4.2.1: *Demographic data for the volunteers*

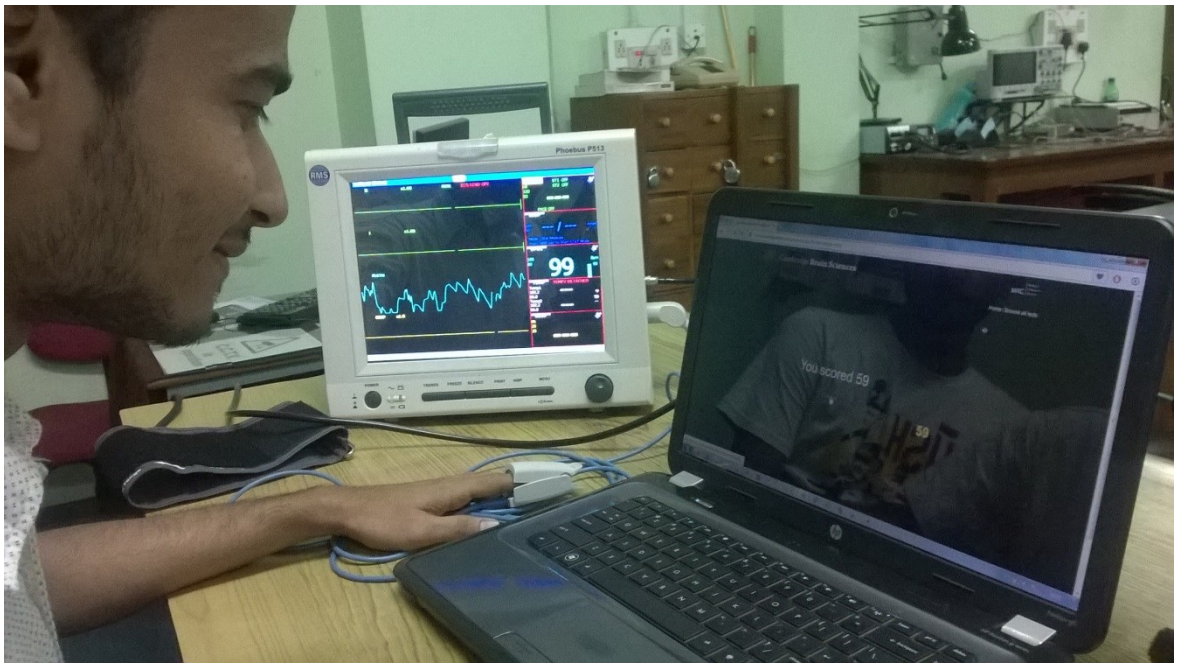


Fig 4.2.1: *Stroop Test Set-up*

4.2.3: Procedure: All experiments were conducted in a quiet room with ambient temperature (23 ± 1°C). The experiments were performed between 9.30 and 12.00 a.m. to minimize circadian variation. All participants attended to this research for two times, firstly, to measure the autonomic nervous system (Heart Rate, Skin Temperature) in absence of Velvet Rose (floral) Aroma and secondly, in presence of Velvet Rose (floral) Aroma while conducting Stroop Test. Before ANS measurement beginning, the researcher clearly informed the procedure, and then participants signed an Informed Consent Form describing the present study and their rights. Heart rate and Skin Temperature were recorded every 30 seconds interval. The test consisted of two sessions. Each session took 3 minutes each. In the second session, the Velvet Rose air-wick was used in room.

4.2.4: Results: The mean and Standard Deviation (SD) values of autonomic parameters in the experiment are presented in Table 4.2.2. The data were compared on various autonomic parameters during presence and absence of floral (Velvet Rose) Aroma. In presence of floral (Velvet Rose) Aroma subjects had significantly reduced Heart rate (six had reduced Heart rate), the average score increased by 9.78%, the body skin temperature showed a non-conclusive pattern, dropping by 1 degree Fahrenheit for some and increasing for others. Though the test subject size is quite small, because of time and monetary constraints, the results attained were in line with those predicted by others.

	Parameter	sb1	sb4	sb5	sb7	sb8	sb9	sb10	sb11
With Aroma	Temp Mean	96.02	93.92	95.93	94.83	95.48	94.32	95.97	96.13
	Temp Std Dev	0.53	0.30	0.46	0.27	0.55	0.42	0.48	0.31
	HR Mean	80.50	92.33	73.33	74.67	76.17	109.67	91.50	89.00
	HR Std Dev	3.27	3.14	2.94	7.45	2.48	8.98	7.23	5.73
	Score Mean	65.00	43.50	11.50	29.00	32.00	22.00	19.00	40.50
Without aroma	Temp Mean	95.83	96.03	96.25	93.05	94.42	95.72	94.23	95.95
	Temp Std Dev	0.16	0.32	0.93	0.46	0.21	0.60	0.50	0.54
	HR Mean	81.83	110.17	82.50	94.67	72.50	99.50	104.67	92.50
	HR Std Dev	6.01	4.45	4.14	12.11	3.73	8.12	2.50	3.45
	Score Mean	63.50	38.00	10.00	23.00	36.00	20.00	15.50	41.00

Table 4.2.2: *Mean and Standard Deviation of ANS during Stroop Test*

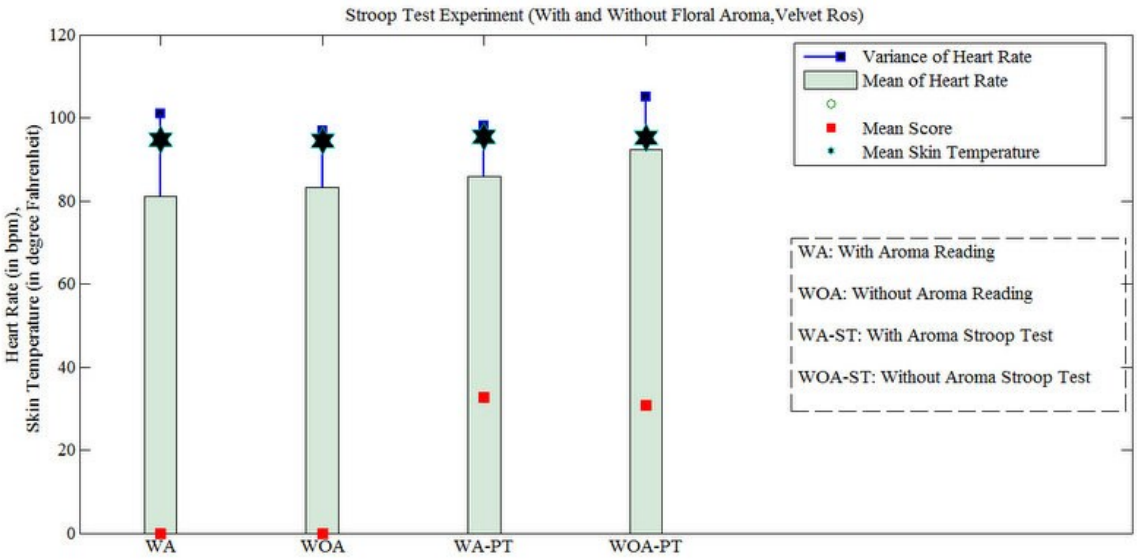


Fig 4.2.2: *Results for Stroop Test*

4.2.5: Conclusions: The presence of Floral (Velvet Rose Aroma), tends to lower the anxiety of subject, increase concentration and alertness. Stroop Test is one way to test the concentration of Person and with time limits imposed; it also tests the alertness level of the person. In [54], the author conducted an experiment that showed that the rosemary group, showed decreased frontal alpha and beta power, suggesting increased alertness. They also had lower state anxiety scores, reported feeling more relaxed and alert and they were only faster, not more accurate, at completing the math computations after the aromatherapy session. The sample of test subjects tested was relatively small (n= 8). Therefore at this stage it would be impossible to generalize the results. The experiment needs to be conducted with a larger sample size

- Six out of eight participants showed a lower hear Rate while performing the Stroop Test in presence of Velvet Rose.
- The Average Score increased by 9.75% in presence of Velvet Rose Aroma.
- The participants' skin temperature showed a non-conclusive Pattern.

4.3: Physical Performance Experiment: In [29], "The effects of peppermint on exercise performance", the author conducted experiment on change in physical performance due to peppermint odour. The results of the experiment support the effectiveness of peppermint essential oil on the exercise performance. Gas analysis, spirometer parameters, blood pressure, and respiratory rate was analysed in the young male students. Relaxation of bronchial smooth muscles, increase in the ventilation and brain oxygen concentration, and decrease in the blood lactate level were offered as the most plausible explanations.

In [52], "Musical Components and Styles Preferred by Young Adults for Aerobic Fitness Activities", the author conducted a study on the attitudes of young adults concerning the influence of musical structural components and style on motor activity. Information was obtained through interviews of 70 college students (35 males and 35 females) enrolled in an aerobic dance class. Respondents (97%) indicated that music made a difference in their class performance. Specifically, musical style (96%), tempo (96%), rhythm (94%), and extra musical associations evoked by music (93%) were the musical components most effective in aiding aerobic activity. Ninety-seven percent of the subjects responded that music improves mental attitude toward the activity, while 79% indicated that music aids in pacing, strength, and endurance. Rock, pop, and new wave music were identified across age subgroups (from ages 18 to 30 years) and by both males and females as the three most frequently preferred musical styles for use in an aerobic workout.

In [53], "Psychophysical and ergogenic effects of synchronous music during treadmill walking", the author analysed the effect of Rock and Pop music on endurance and a range of psychophysical indices during a treadmill walking task. The results indicated that motivational synchronous music can elicit an ergogenic effect and enhance in-task affect during an exhaustive endurance task.

4.3.1: Physical Performance Test: Subjects were asked to climb the stairs to first floor of departmental building, twice to the best of speed possible. Heart Rate, Respiration Rate, Skin Temperature and Blood Pressure were measured to analyse the effect of Peppermint on their performance.

4.3.2: Test Subject: Eight selected participants all male aged between 18 and 24 years (mean age 20.5 ± 1.67 years) with normal body mass indices (mean 24.89 + 3.59) were enrolled in the present study. A summary of the demographic data of the participants is presented in Table 4.2.1. None of the subjects had abnormalities affecting smell, cardiovascular diseases, or a history of smoking or drug addiction. Twelve hours prior to testing subjects were asked to wash their hair without any spray. They were also asked not to use antiperspirants, perfumes and refrain from consuming alcohol, cigarettes, or caffeinated drinks. They were requested to try to sleep well before the day of the experiment to avoid feeling fatigued or drowsy. Subjects were given a full explanation of the research and a written informed consent of all aspects of the present study, and were free to withdraw at any time.

Parameter	N	Minimum	Maximum	Mean	SD
Age (yrs)	8	18	23	20.5	1.41
Height (cm)	8	160	186	178	6.71
Weight (kgs)	8	68	102	78	10.99
Body Mass Index	8	21.39	31.48	24.61	3.08

Table 4.3.1: *Demographic data for the volunteers*

4.3.3: Procedure: The experiment was repeated twice in the evening 4:00 to 5:30PM on two consecutive days, once when the subjects were exposed to peppermint aroma and secondly, without any aroma. Before ANS measurement beginning, the researcher clearly informed the procedure, and then participants signed an Informed Consent Form describing the present study and their rights. Heart rate and Skin Temperature were recorded every 30 seconds interval. The test consisted of two sessions. Each session took 10 minutes each. In the first session, the subject was asked to inhale peppermint essential oil (diluted with water to 10% V/V), for half an hour before beginning the physical activity.

4.3.4: Results: The mean and Standard Deviation (SD) values of autonomic parameters in the experiment are presented in Table 4.3.2. The data were compared on various autonomic parameters during presence and absence of Peppermint Aroma. In presence of Peppermint Aroma subjects had significantly reduced Heart rate (six out of eight had reduced Heart rate), Systolic Pressure increased for six out of eight participants, while diastolic pressure reduced for seven out of eight participants. The breathing rate reduced for four participants, remained same for two and increased for other two participants. The skin temperature followed a more haphazard trend and was inconclusive because of small test subject size. Though the test subject size is quite small, because of time and monetary constraints, the results attained were in line with those predicted by others.

	Parameter	sb1	sb2	sb3	sb4	sb5	sb6	sb7	sb8
With Peppermint	HR Mean	90.33	100.67	125.67	75.33	95.00	87.00	93.67	112.33
	HR Std. Dev.	1.53	4.93	6.11	6.66	6.08	3.61	27.79	18.90
	BP Systolic Mean	178.00	150.00	192.67	157.33	161.33	177.33	177.67	182.00
	BP Systolic Std. Dev.	3.61	1.73	4.51	10.79	22.23	11.85	15.01	8.89
	BP Diastolic Mean	72.67	66.67	75.00	69.33	68.33	66.67	96.33	80.33
	BP Diastolic Std. Dev.	1.15	6.51	2.00	5.13	4.51	1.15	4.16	5.03
	BR Mean	21.33	20.00	28.67	27.00	26.00	29.33	18.67	24.00
	BR Std. Dev.	3.06	0.00	3.06	3.00	2.00	2.31	1.15	4.00
	Temp Mean	93.83	96.60	94.73	97.20	95.43	94.37	95.17	97.00
	Temp Std. Dev.	0.31	0.62	0.90	0.20	1.68	0.42	1.27	0.46
Without Peppermint	HR Mean	117.67	103.00	126.00	76.67	70.00	103.67	83.67	117.67
	HR Std. Dev.	4.04	8.54	5.20	2.08	12.17	6.11	30.89	4.04
	BP Systolic Mean	162.33	146.00	190.00	168.33	153.67	167.00	182.67	162.33
	BP Systolic Std. Dev.	12.74	5.57	1.00	10.69	3.06	8.72	9.50	12.74
	BP Diastolic Mean	88.33	74.33	76.67	77.00	75.00	73.67	75.67	88.33
	BP Diastolic Std. Dev.	1.15	1.53	2.08	2.65	5.29	3.51	1.53	1.15
	BR Mean	24.00	22.00	31.33	28.67	20.67	29.33	17.33	24.00
	BR Std. Dev.	2.00	2.00	3.06	5.03	1.15	3.06	2.31	2.00
	Temp Mean	94.37	95.67	95.60	96.73	94.83	95.50	96.83	94.37
	Temp Std. Dev.	0.42	0.21	0.20	0.31	0.25	0.36	0.15	0.42

Table 4.3.2: *Mean and Standard Deviation of ANS during Physical Performance Test*

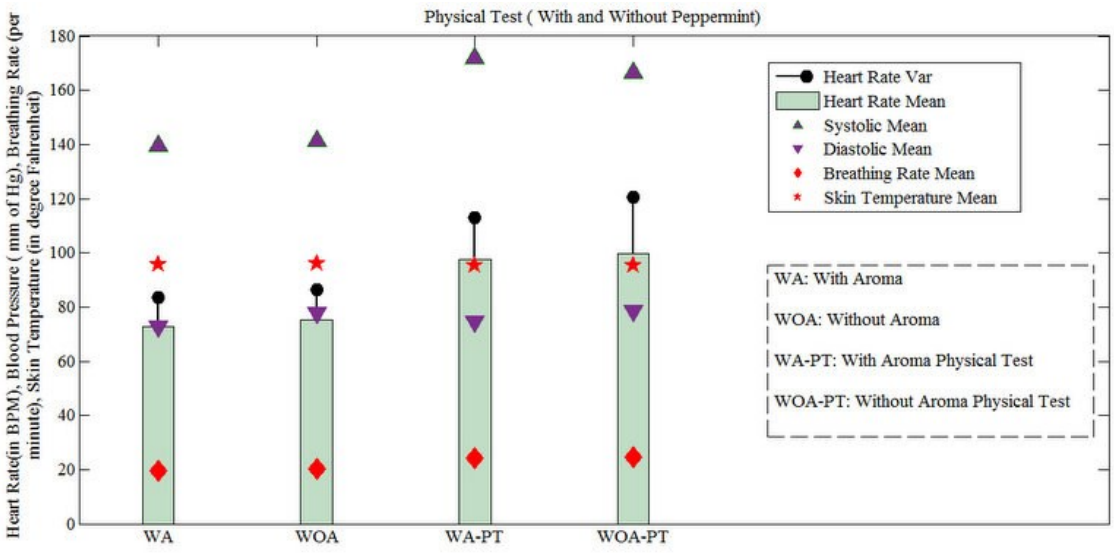


Fig: 4.3.1: *Results for Physical Test*

4.3.5: Conclusions: The results of the experiment support the effectiveness of peppermint essential oil on the physical performance. In presence of peppermint, the participants showed a lower breathing Rate and Heart Rate while doing the same work. The sample of test subjects tested was relatively small (n= 8). Therefore at this stage it would be impossible to generalize the results. The experiment needs to be conducted with a larger sample size.

- The Heart Rate dropped by 2.29% on an average in presence of peppermint.
- The Systolic blood pressure increased by 3.30%, whereas Diastolic blood pressure decreased by 5.35% in presence of peppermint.
- The breathing rate also decreased by 4.00% in presence of peppermint

Chapter 5: Algorithm Development and System Description

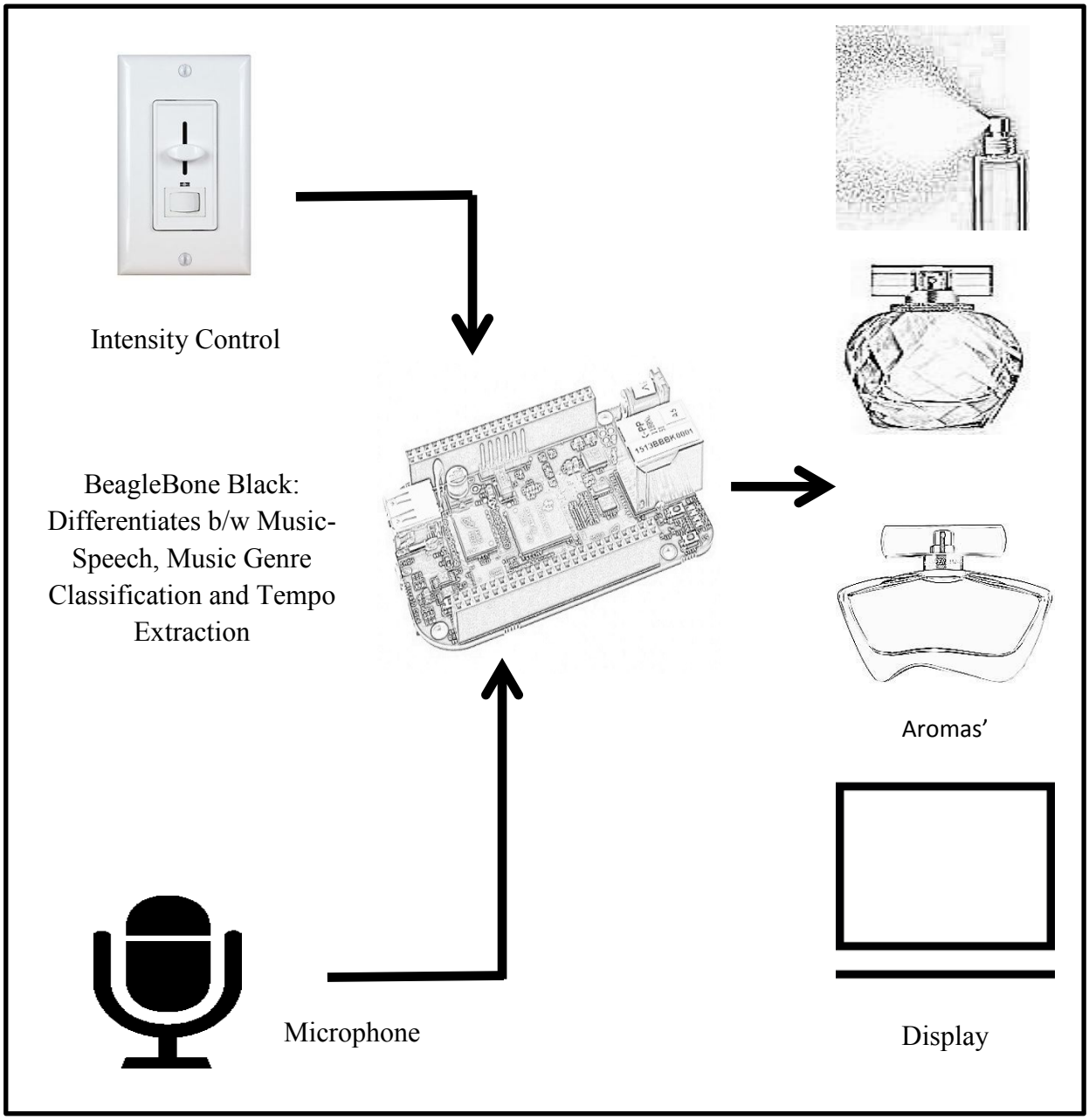


Fig 5.1.1: *Component Description of System*

5.1: Components

Beagle Bone Black:

- Processor – AM335x 1GHz ARM ® Cortex-A8
- 512MB DDR3 RAM
- 4GB 8-bit eMMC on-board flash storage
- 3D Graphics accelerator
- NEON floating-point accelerator
- 2x PRU 32-bit microcontrollers
- 2X46 headers for expansion and cape connectivity
- USB client for Power & Communication
- USB Host
- Ethernet
- HDMI

Microphone: Genius MIC-02A Microphone

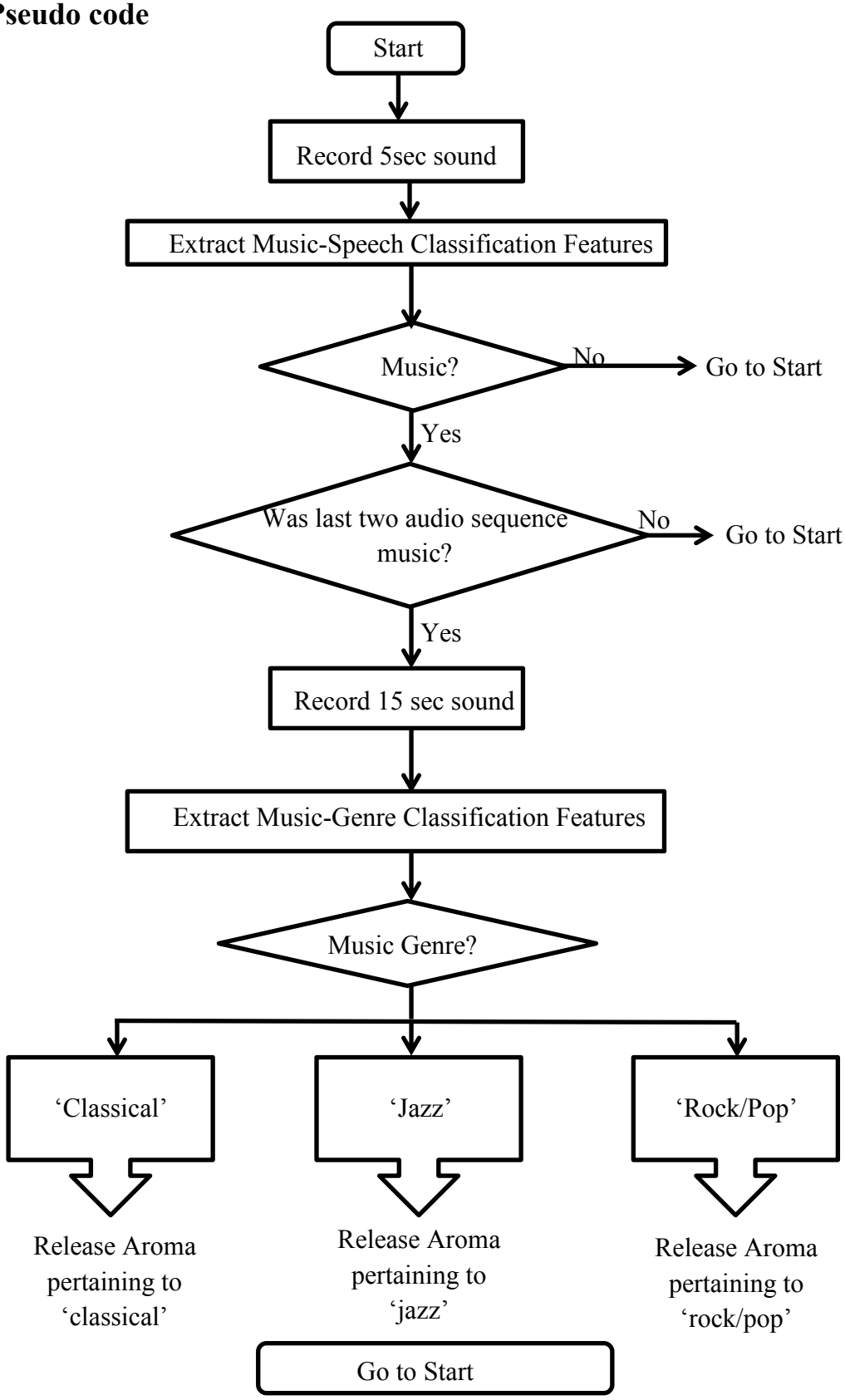
- Omni-Directional Microphone
- 100 Hz - 10 KHz Frequency Response
- -42dB+/-3dB Sensitivity
- 2.2kohm Impedance

Sound Card: Ad-Net USB Sound Card 3D Virtual 5.1

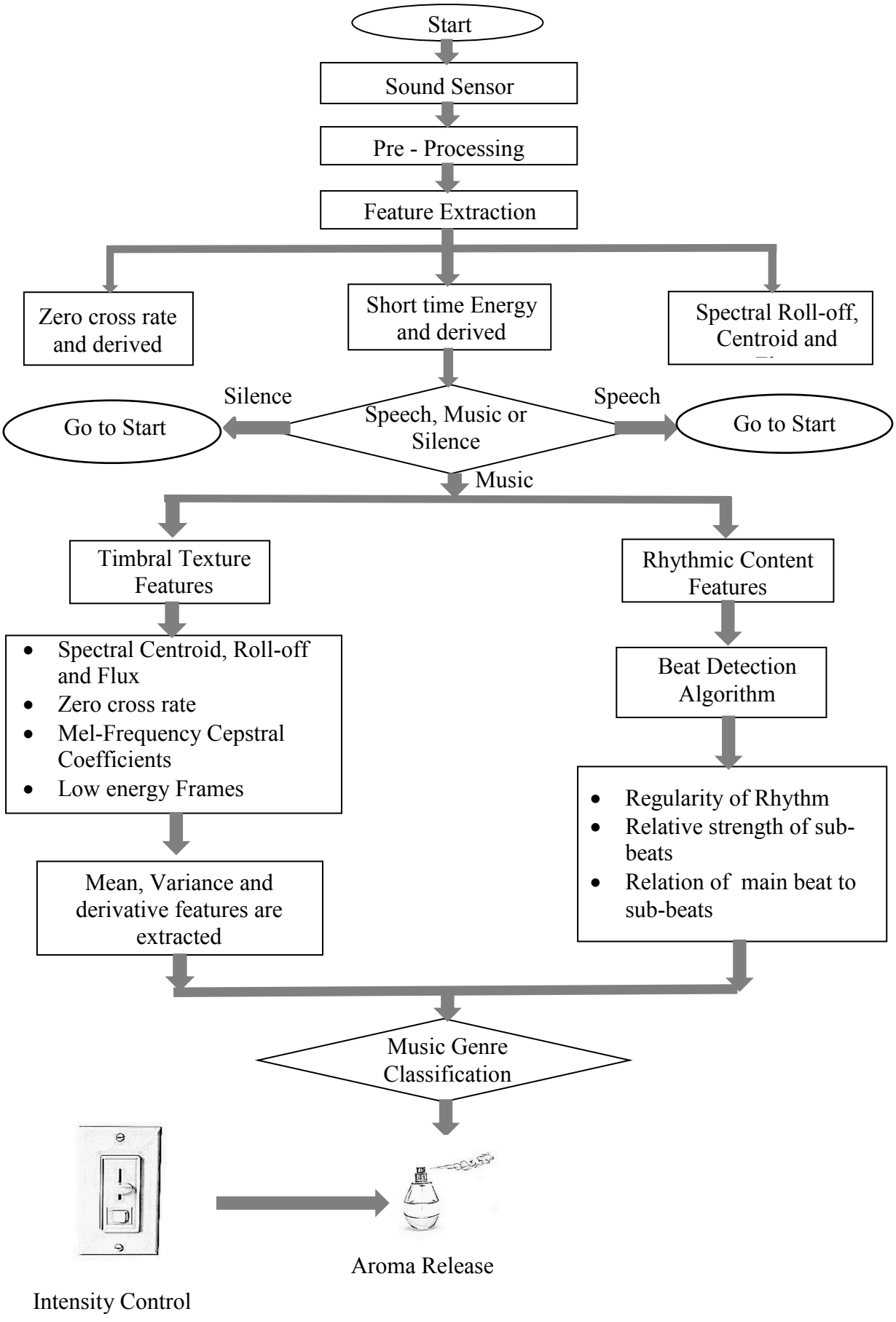
- Use USB port power
- Support 3D positional sound and virtual 5.1 CH sound track
- Digital Class-B Power Amplifier inside
- No-External Power required

Intensity Control: Potentiometer (0-10 k-ohm)

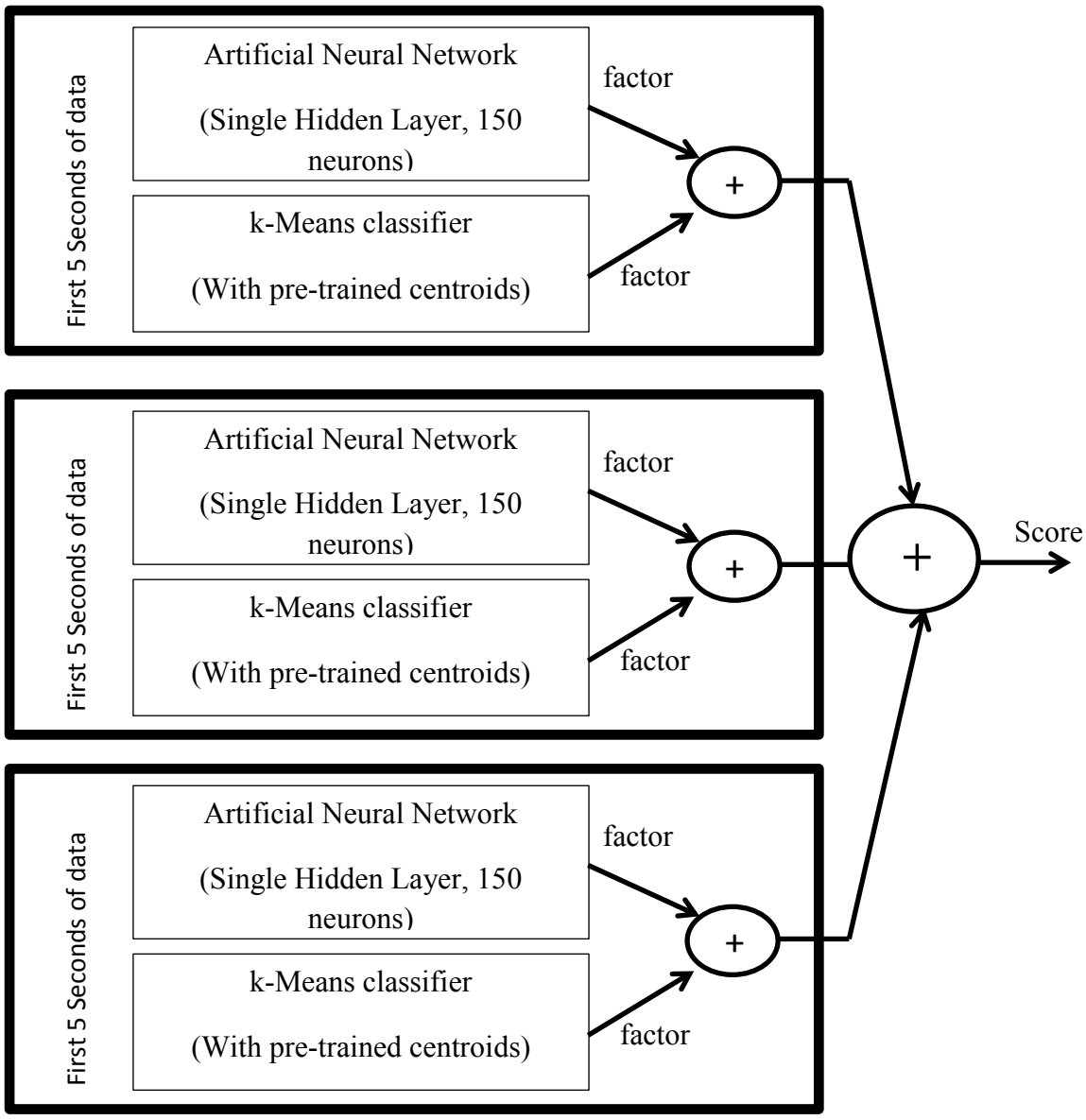
5.2: Pseudo code

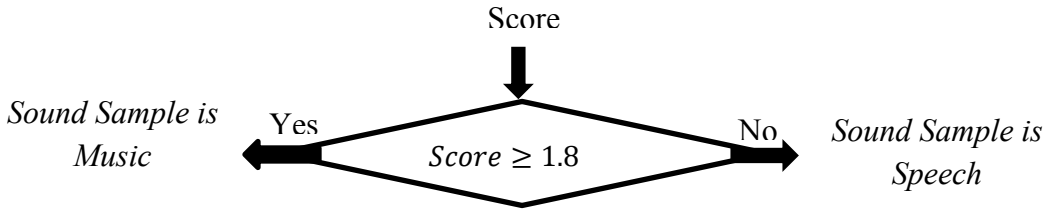


Flowchart 5.2.1: *Pseudo Code*



Flowchart 5.2.2: *Feature Extraction flow*





Flowchart 5.2.3: *Music-Speech Classifier*

	Music	Speech
ANN	0.6	0.4
k-Means	0.4	0.6

Table 5.2.1: *Values of factor*

5.3 Aroma Release Circuit

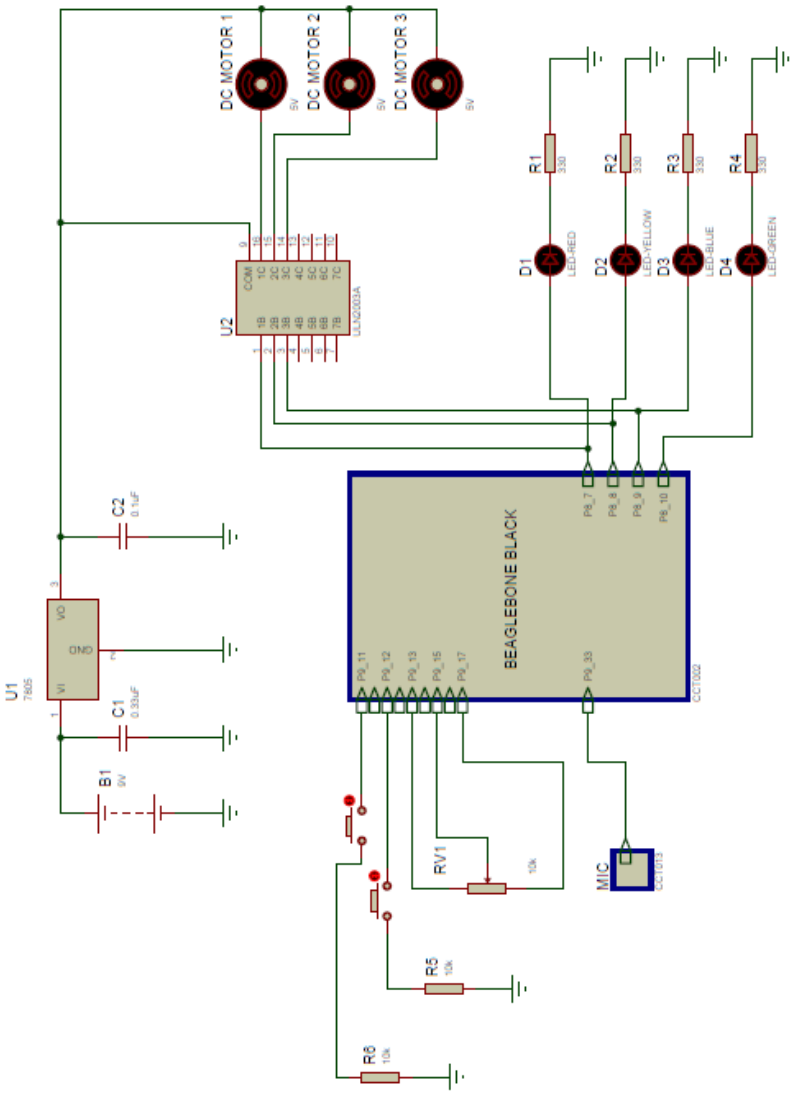


Fig 5.3.1: *Aroma Release Circuit Diagram*

Beaglebone Black collects the audio signal from environment using microphone, processes it and based on the music genre and tempo; use the information to release a specific Aroma. Air-Wick modules were used for Aroma Release. IC ULN2003A, was used to drive the Air-wick DC motors. The IC ULN2003A is high-voltage, high-current Darlington transistor arrays. Each consists of seven NPN Darlington pairs that feature high-voltage outputs with common-cathode clamp diodes for switching inductive loads. The collector-current rating of a single Darlington pair is 500 mA.

5.4: Aroma Release Pseudo Code

Step 1: Sample the value of Potentiometer (10 kΩ). Depending on the value of potentiometer, the intensity of aroma will be controlled.

Step 2: Record 15 seconds Audio signal at 8kHz; break it down in three 5 seconds Audio samples.

Step 3: For each 5 second audio, extract Music-Speech Classification features (Short Time Energy, Short Time Zero Cross Rate, Spectral Roll Off, Spectral Flux and Spectral Centroid).

Step 4: Using Ensemble Classification (k-Means and ANN classifiers), classify the audio signal to Speech or Music.

Step 5: If, the classified signal is Speech, then go back to Step2. Else, extract Music-Genre classification features (Short Time Energy, Short Time Zero Cross Rate, Spectral Roll Off, Spectral Flux, Spectral Centroid, MFCC and Spectral Contrast)

Step 6: Using Ensemble Classification (Multiple ANN Classifiers), classify the recorded Audio in one of the Genres (Classical, Jazz or Pop/Rock).

Step 7: Record another 15 seconds Audio Signal and reclassify it to a specific Genre.

Step 8: After t time (t is selected based on potentiometer reading), release the Aroma pertaining to the genre which was classified most number of times.

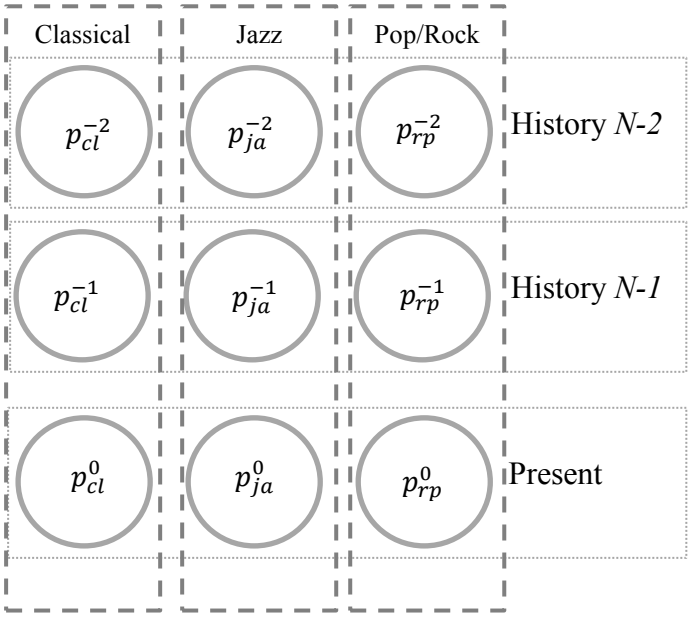
Step 9: Next, the process will be repeated from Step 1, but this time we will also consider history to release the Aroma. Only, last two released Aroma's will be considered. The algorithm to do is explained below.

$$P_{cl} = 0.5 \times p_{cl}^{0} + 0.3 \times p_{cl}^{-1} + 0.2 \times p_{cl}^{-2} \qquad \boldsymbol{Eq.5.1}$$

$$P_{ja} = 0.5 \times p_{ja}^{0} + 0.3 \times p_{ja}^{-1} + 0.2 \times p_{ja}^{-2} \qquad \boldsymbol{Eq.5.2}$$

$$P_{rp} = 0.5 \times p_{rp}^{0} + 0.3 \times p_{rp}^{-1} + 0.2 \times p_{rp}^{-2} \qquad \boldsymbol{Eq.5.3}$$

:



The history is considered only after two Aroma release. This is done in order to ensure that, the transition from one Aroma to another is smooth and not abrupt. Also, the boundaries between two genres are not well defined, and the human genre specification might vary person to person. The history smoothens out effect of such variations.

Let's suppose for first t time (here t is controlled by value of potentiometer), total N classifications of Music Genre were carried out; then probabilities p_x,(where *x* defines the genre) are calculated by maximum likelihood estimation.

$$p_{cl} = \frac{N_{cl}}{N} \qquad \boldsymbol{Eq.5.4}$$

$$p_{ja} = \frac{N_{ja}}{N} \qquad \boldsymbol{Eq.5.5}$$

$$p_{ro} = \frac{N_{ro}}{N} \qquad \boldsymbol{Eq.5.6}$$

These are normalized by dividing maximum of the $\{p_{cl}, p_{ja}, p_{ro}\}$. The history of last two t minutes window is considered to release the current aroma, based on max of ($Eq.5.1$, $Eq.5.2, Eq.5.3$)

5.5 Music based Aroma Release System

Fig 5.5.1: *Aroma release System*

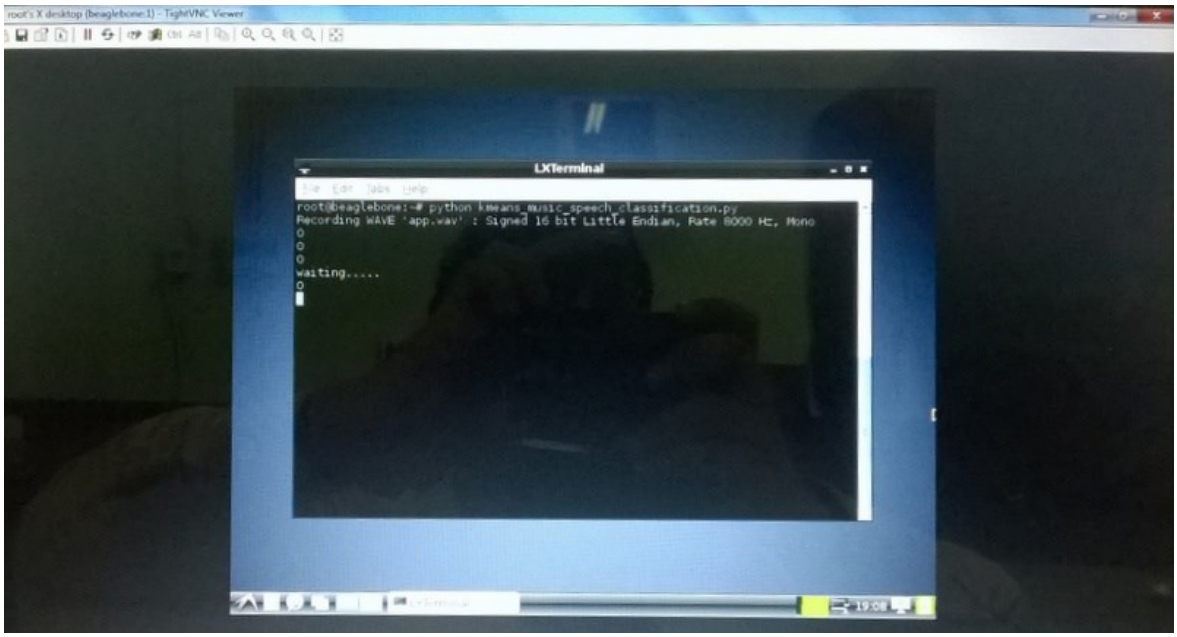


Fig 5.5.2: *Beaglebone terminal*

Chapter 6: Results

Part A- Matlab Results

6.1: Music-Speech Classification

	Mixed Music	Blues	Classical	Country	Disco	Hip-hop	Jazz	Metal	Pop	Reggae	Rock
All features	89.35	94.43	89.38	94.95	94.98	89.36	95.52	100	93.30	91.05	91.60
Zero Cross Features	87.16	90.51	94.97	88.28	74.31	82.61	94.38	91.03	85.49	76.52	80.98
Energy Features	92.76	91.05	83.77	91.05	96.63	92.16	91.98	98.87	97.77	90.18	93.84
Spectral Features	94.42	89.96	90.49	90.48	91.59	84.91	94.96	99.44	83.24	79.89	86.56

Table 6.1.1: *Results for ANN for various combinations of Feature sets and Music Genre*

Energy and Spectral Feature set provide good discrimination for Music-Speech classifier using an ANN in Matlab. The .wav sounds used were noise free and 30 seconds in length. Zero cross rate based feature set usually under-performs in case of Disco and Reggae music genre.

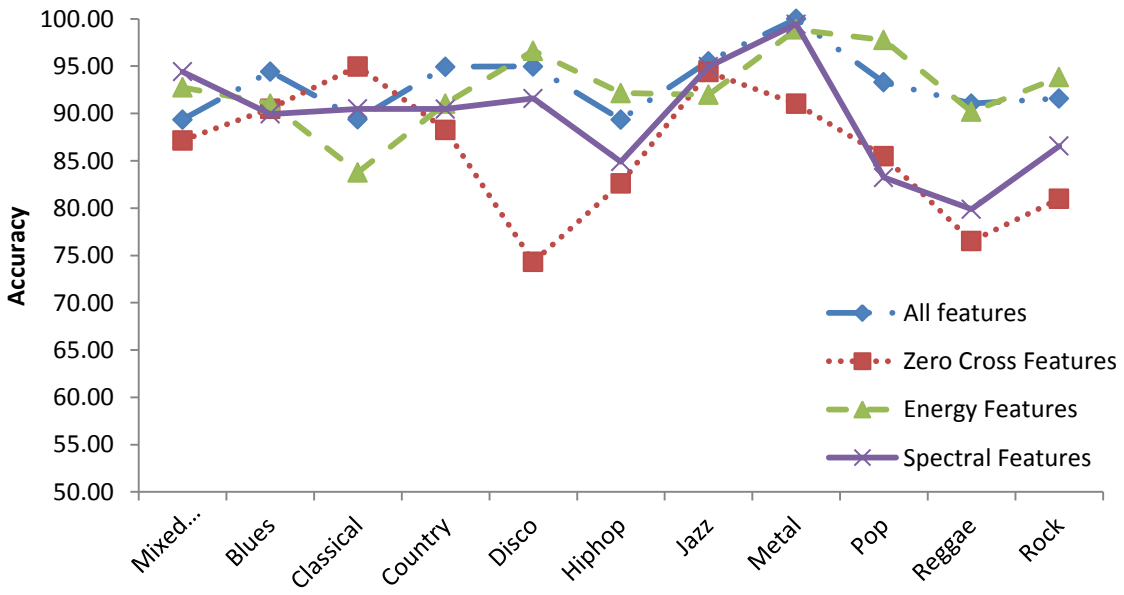


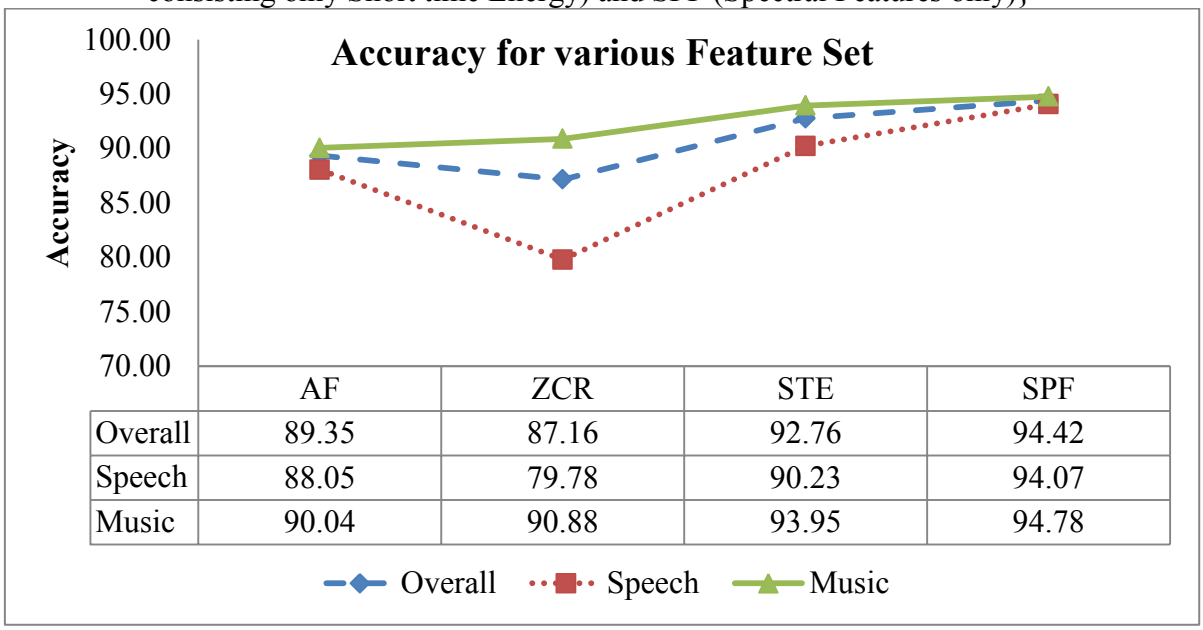
Fig 6.1.1: *Music Speech Accuracy plot for various genres*

All Features				Energy Features			
% data for training	Overall Accuracy	Speech accuracy	Music accuracy	% data for training	Accuracy overall	Speech accuracy	Music accuracy
50	81.25	78.95	82.22	**50**	93.75	89.47	95.56
60	94.11	89.47	96.88	**60**	92.15	94.74	90.63
70	94.73	100.00	92.00	**70**	97.36	92.31	100.00
80	92.00	90.00	93.00	**80**	84.61	80.00	87.50
Zero cross Rate Features				**Spectral Features**			
% data for training	Accuracy overall	Speech accuracy	Music accuracy	% data for training	Accuracy overall	Speech accuracy	Music accuracy
50	81.25	68.42	86.67	**50**	98.43	100.00	97.78
60	92.15	89.47	93.75	**60**	92.16	100.00	87.50
70	89.47	92.31	88.00	**70**	94.73	92.31	96.00
80	88.46	70	100	**80**	88.46	70.00	100.00

Table 6.1.2: *Results for ANN for various combinations of Feature sets and Mixed Music*

In this case the music dataset used comprised of examples taken from all the different music genres (total ten as mentioned in table D1.1). There is little difference between performance of Energy, Spectral and Zero cross rate features, but Energy does seem to offer a higher accuracy.

{AF (all features), ZCR (feature vector consisting only zero cross rate), STE (feature vector consisting only Short time Energy) and SPF (Spectral Features only)}



	AF	ZCR	STE	SPF
Overall	89.35	87.16	92.76	94.42
Speech	88.05	79.78	90.23	94.07
Music	90.04	90.88	93.95	94.78

Fig 6.1.2: *Accuracy plot for mixed music for various feature set*

The accuracy attained for Music was more than that for speech in all the cases.

6.2: Music Genre Classification

For Music Genre classification, two ANN based structures were tried. First is single stage structure, and second is two stages, two ANN structure.

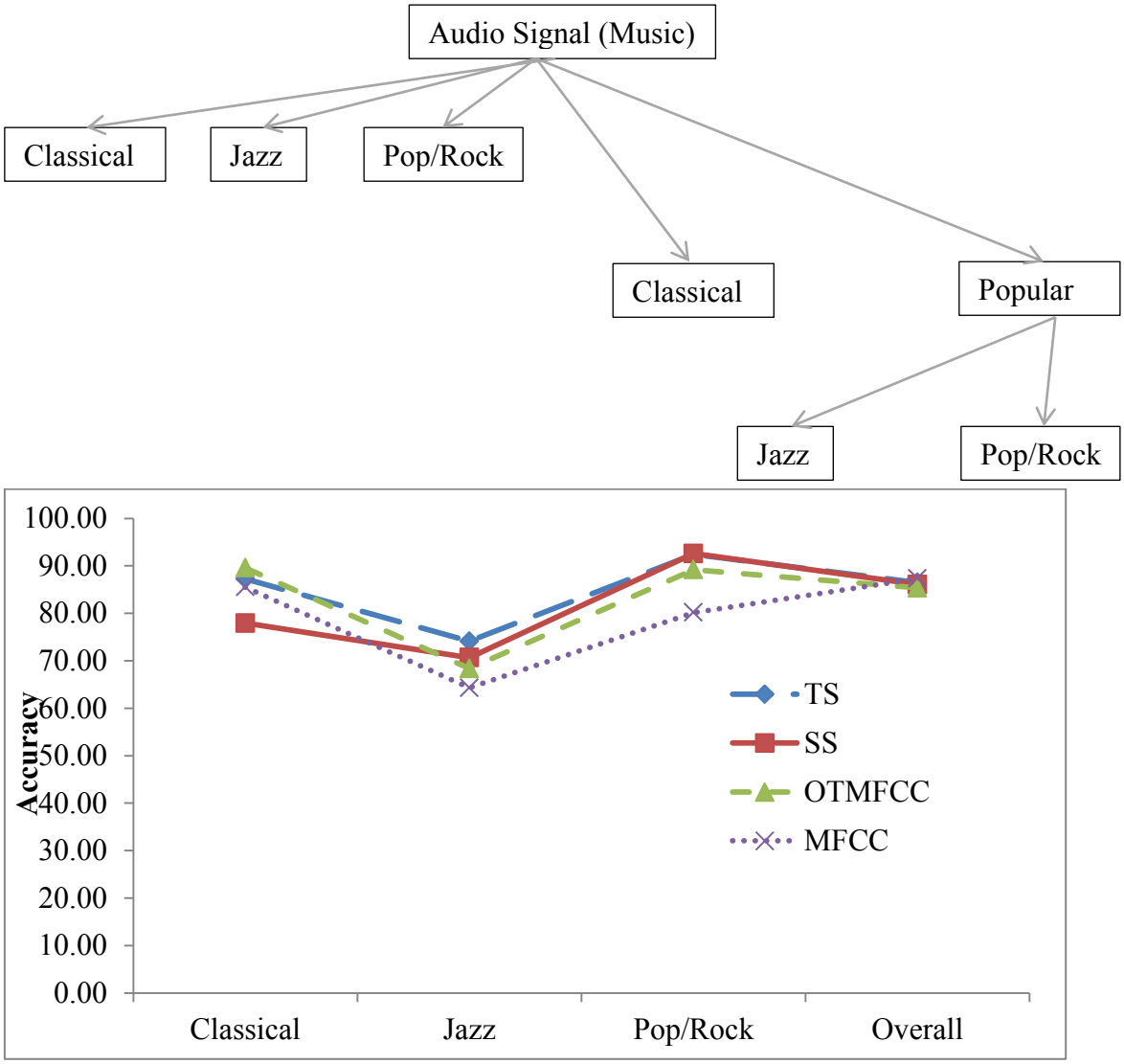


Fig 6.2.1: *Accuracy for various feature set size and Prediction Structures.*

(TS: Two Stage; SS: Single Stage; OTMFCC: Other than MFCC feature Set; MFCC: MFCC based feature set)

	Classical	Jazz	Pop/Rock	Overall
Two Stage	87.24	74.12	92.41	86.49
Single Stage	77.93	70.66	92.60	86.07
Other than MFCC	89.58	68.37	89.20	85.36
MFCC Based feature set	85.50	64.29	80.14	87.32

Table 6.2.1: *Music genre prediction Accuracy.*

6.3: Tempo Classification

BPM	Speed of Music
<60	Very Slow
<100	Slow
100-160	Medium
160-240	Medium Fast
>240	Fast
>320	Very Fast

Table 6.3.1: *BPM to Speed Conversion table*

Beat Algorithm Accuracy:

- 75% correct BPM
- 15% Half or Double
- 10% Incorrect

Dataset: Giansteps tempo Dataset, focus on the Electronic Dance Music. It can be downloaded from https://github.com/GiantSteps/giantsteps-tempo-dataset . The data sets have been automatically created from user feedback and annotations extracted from web sources. There is a consensus in the tempo-tracking literature on the fact that state-of-the-art tempo induction algorithms typically make errors of metrical levels (they output e.g. half the correct tempo). Accordingly, the tempo-tracking algorithm we use here DWT may output the correct tempo, twice or half of its value or two thirds of its tempo.

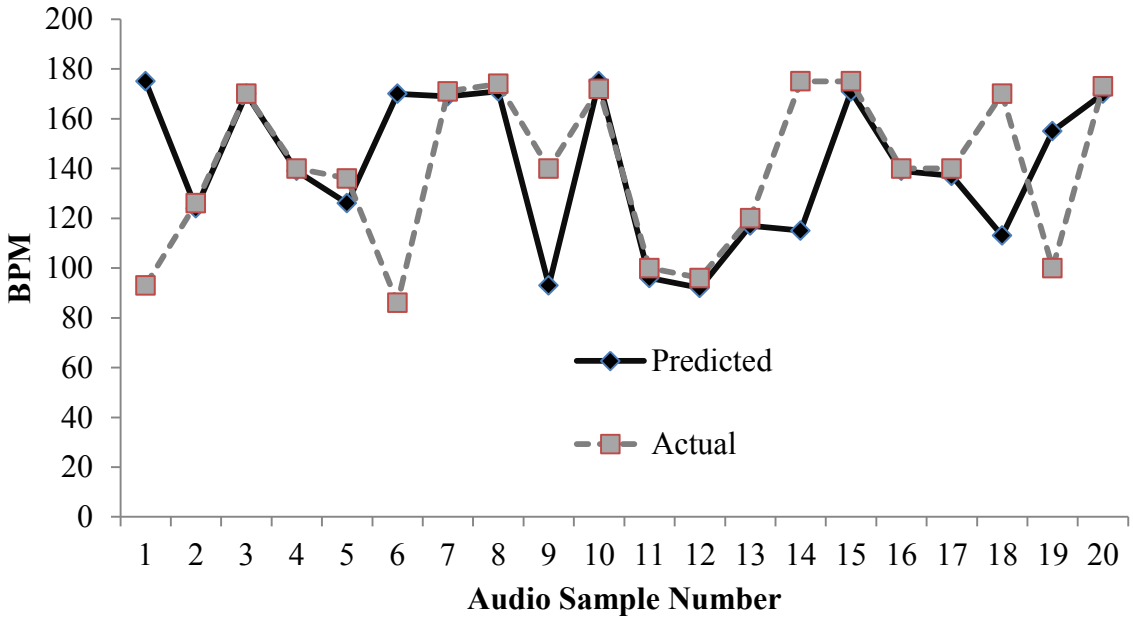


Fig 6.3.1: *Calculated BPMs' for Electronic Dance Music*

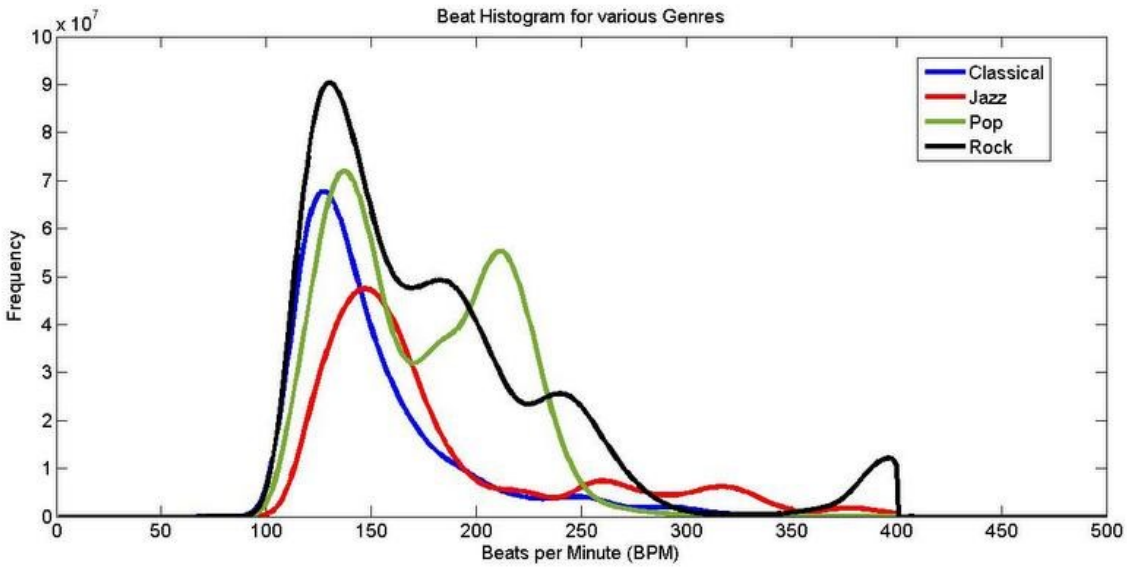


Fig 6.3.2: *BPM histogram for various genres (20 songs of each genr*e)

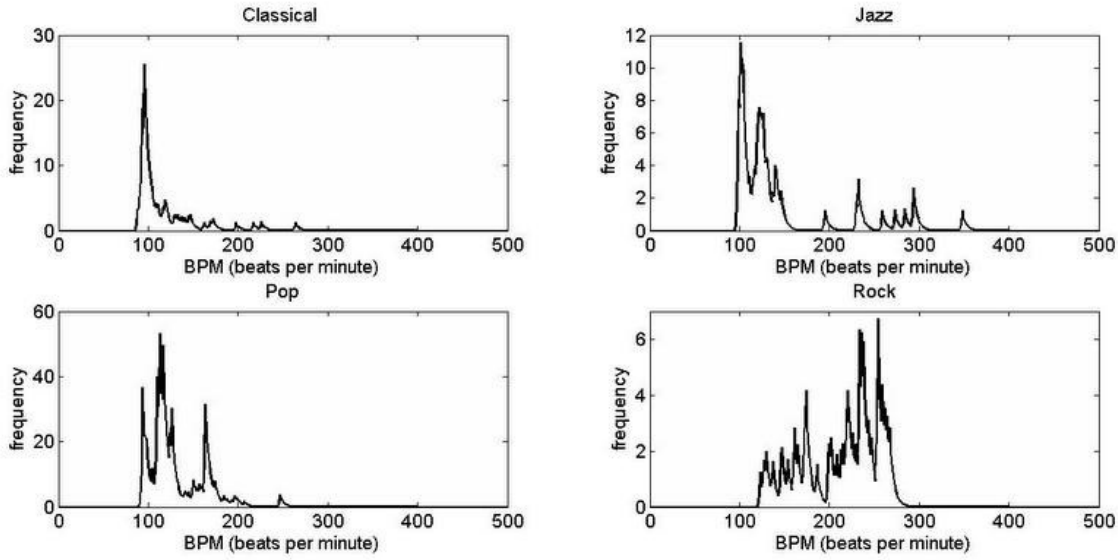


Fig 6.3.3: *Beat histograms for an individual Audio samples*

As can be observed in Fig. 6.3.3, Classical has most consistency in BPM, (i.e. the BPM varies least for classical Songs). Jazz, though have concentrated BPM in 100 to 180BPM range, it also has more than one BPM with small differences. Pop and Rock genres can have quite a wide range of BPM's and can have sudden shifts. Some authors, [57] used BPM histogram information for Music Genre classification.

Part B – Embedded System Results

6.4: Music Speech Classification:

I. The data for training was also recorded using the embedded system with active human participation. Each recording was of 5 seconds.

II. For each audio-file two to three samples was collected for training. Overall 60% of the dataset was used for training ANN and k-Means algorithm.

III. FANN, (A fast implementation of ANN in C++ was used using pyfann library for Debian armhf architecture.)

ANN Classifier		
	Music	**Speech**
Music	93.75	6.25
Speech	20.32	79.68

k-Means Classifier		
	Music	**Speech**
Music	81.25	18.75
Speech	6.25	93.75

Table 6.4.1: *ANN & k-Means accuracy*

IV. Artificial Neural networks performs better for Music with an accuracy of 93.75% compared to just 79.68% for Speech.

V. K-Means Classifier performs better for Speech with an accuracy of 93.75% compared to just 81.25% for Music.

VI. To improve the overall accuracy we used a 15 second recording, which was later broken in three 5 seconds window and analysed for Music-Speech Classification. The results showed an improved accuracy in both Music, as well as Speech Classification.

VII. The 'Ensemble Learning' was used to increase the accuracy, where both ANN and k-Means had equal weight and if 4out of 6 predictions was of Music, it was classified as Music, else Speech.

k-Means + ANN Classifier		
	Music	**Speech**
Music	84.37	15.63
Speech	9.37	90.63

Table 6.4.2: *ANN + k-Means accuracy*

VIII. The 'Speech with continuous zzz…… sound', was often misclassified as Music.

IX. The 'songs with rap' were also often misclassified as Speech.

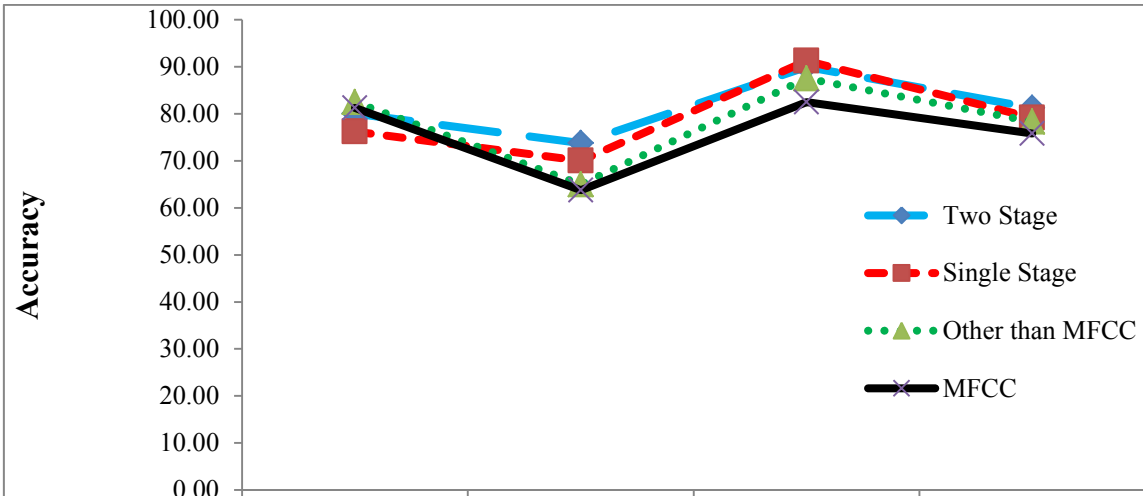
6.5: Music Speech Classification:

The embedded system was trained using YouTube Audio Library https://www.youtube.com/audiolibrary/music . It consists of music files sorted based on genre, mood, Instrument, duration and Attribution. The embedded system was trained on 200 audio files of each genre (each audio is 15 seconds in length) and was tested for 80 audio files (each 15 seconds in length).

I. The accuracy observed for various genres turns out to be close to the range achieved using Matlab, but we needed to expand the feature set by including first different of Short Time Energy and Zero Cross Rate.

II. The observed accuracy for various genres is in range of 70-80%, which is better than random allocation (which would be 33.33%).

III. The classifiers based on Two Stage, One Stage, MFCC and Other than MFCC based feature set were used in Ensemble classification to ensure more consistency in results.



	Classical	Jazz	Pop/Rock	Overall
Two Stage	80.00	73.75	90.00	81.25
Single Stage	76.25	70.00	91.25	79.17
Other than MFCC	82.50	65.00	87.50	78.33
MFCC	81.25	63.75	82.50	75.83

Fig 6.4.1: *Music genre prediction accuracy*

{TS = Two Stage Classifier; SS = Single Stage Classifier; OTMFCC = Other than MFCC based feature set Classifier; MFCC = MFCC Based Classifier}

Chapter 7:

Conclusions

The target of the project was to develop an Embedded System to release specific Aroma, at a controlled rate based on genre or tempo of Music. Though there is no clear definition of a Music genre, but based on studies [58], Classical, Jazz, Pop and Rock were selected as the most basic Music Genres. Since no standard literature exist regarding co-relating a music genre with an Aroma, the effects of music genre and Aroma on a test subject were considered as basis for correlating them. In [28], the author proposed that Lavender causes relaxation, and similarly in [40] the author analysed the effect of classical Music on progressive relaxation. Hence, Lavender was correlated with Classical Music. Similarly, in [41], the author proposed the hypothesis based on experiment that Jazz music results in increased Mental performance, and in [34], the author analysed the effect of Velvet Rose on Mental Performance. Hence Jazz music genre was correlated with Velvet Rose. Similarly from studies conducted in [35], [52] and [53] Pop/Rock was correlated with Peppermint Aroma. So, an experiment was designed to analyse the effects of Aroma's on Human's Physical and Mental performance and his/her relaxation. Trier Social Stress Test was conducted to verify the effects of Lavender on Relaxing, Stroop Test to verify effect of Velvet Rose on Concentration and Physical (Stair Climbing) Test to verify peppermint increases physical performance. Though the test subject set was small (N=12), the experiments were repeated (twice in case of Stroop Test and three times in case of Physical Test). The results attained were in line with those reported in literature.

An embedded system with BeagleBone Black was constructed, with microphone (input sensor), Potentiometer (to control Intensity) and three air wick modules (which were used to release the Aroma) [The picture of final system is shown on page no. 44]. The feature set was selected based on Matlab simulation of Music Genre classifier. Spectral Contrast and Beats Per Minute (BPM) Histogram based features were not considered in final Embedded System development, as they did not increase accuracy by much, but increased feature extraction time by 30%. Artificial Neural Networks were used to classify Music Genre. To increase the accuracy, ensemble classification concept was used. The system was tested for five days.

Future Work

The system developed was a prototype model of the actual product. The system is lab ready, but to target it as a consumer product the system needs to be made more portable, and sturdy.

- In current System, the processor is powered by USB Power form a Computer. This limits portability of the system. So one part could be to design and fabricate a Power Distribution Module (PDM), which powers the microcontroller, and other ICs. The PDM must also ensure safe power flow setup to limit damage to the components.
- The set of LEDs', DC Motor Driver IC and Voltage regulator, were all wired together using a bread-board, which is temporary lab set-up. There is a need to design Printed Circuit Board (PCB), to connect all the ICs.
- Third, currently we are using Beaglebone Black as a processor, which has a lot of redundant hardware, so to better utilise the processor and lower power consumption, it would be better to migrate to ARM processors.
- The experiments done in this work were limited to correlating Aroma with Subject physical and mental performance and relaxation. The verification of aroma with music needs to be carried out to analyse the effect of music genre and aroma combined together.
- Similarly, a set of experiments need to be conducted to analyse the effect of Music on Subjects Physical, and Mental performance and his relaxation.
- The system designed cannot update its ANN training data without connecting to a laptop. In future, when targeting the product to a consumer, the system must be able to upgrade its training dataset from time to time using Bluetooth or Wi-Fi module.

A more advanced approach considered was that the system could be able to predict the emotion of a person from his/her speech, get Autonomous Nervous Systems(ANS) readings from a smart watch through Bluetooth or Wi-Fi and using that information release Aroma and play Music to enhance the pleasure of subject and help him relax/calm down and enjoy music. The selection of Music could be from predefined library.

Part B

Linear and Non-Linear global features based classification of Emotional Speech

Chapter 1: Introduction

Emotions are complex. According to some theories, they are a state of feeling that results in physical and psychological changes that influence our behaviour [2]. The physiology of emotion is closely linked to arousal of the nervous system with various states and strengths of arousal relating, apparently, to particular emotions. Emotion is also linked to behavioural tendency. Extroverted people are more likely to be social and express their emotions, while introverted people are more likely to be more socially withdrawn and conceal their emotions. Emotion is often the driving force behind motivation, positive or negative. An alternative definition of emotion is a "positive or negative experience that is associated with a particular pattern of physiological activity" [4]. According to other theories, emotions are not causal forces but simply syndromes of components, which might include motivation, feeling, behaviour, and physiological changes, but none of these components is the emotion. Nor is the emotion an entity that causes these components.

Emotion is, in everyday speech, a person's state of feeling in the sense of an affect [1], [2]. Emotion is often intertwined with mood, temperament, personality, disposition, and motivation. On some theories, cognition is an important aspect of emotion. Those acting primarily on emotion may seem as if they are not thinking, but mental processes are still essential, particularly in the interpretation of events. For example, the realization of danger and subsequent arousal of the nervous system (e.g. rapid heartbeat and breathing, sweating, muscle tension) is integral to the experience of fear. Other theories, however, claim that emotion is separate from and can precede cognition.

1.1: Classification of Emotions

A distinction can be made between emotional episodes and emotional dispositions. Emotional dispositions are also comparable to character traits, where someone may be said to be generally disposed to experience certain emotions. For example, an irritable person is generally disposed to feel irritation more easily or quickly than others do. Finally, some theorists place emotions within a more general category of "affective states" where affective states can also include emotion-related phenomena such as pleasure and pain, motivational states (for example, hunger or curiosity), moods, dispositions and traits [5].

The classification of emotions has mainly been researched from two fundamental viewpoints. The first viewpoint is that emotions are discrete and fundamentally different constructs while the second viewpoint asserts that emotions can be characterized on a dimensional basis in groupings. In this report the emotions are believed to be fundamentally different and hence all work makes this basic assumption.

1.2: Basic Emotions: Paul Ekman has supported the view that emotions are discrete, measurable, and physiologically distinct. Ekman's most influential work revolved around the finding that certain emotions appeared to be universally recognized, even in cultures that were preliterate and could not have learned associations for facial expressions through media. Another classic study found that when participants contorted their facial muscles into distinct facial expressions (e.g. disgust), they reported subjective and physiological experiences that matched the distinct facial expressions. His research findings led him to classify six emotions as basic: anger, disgust, fear, happiness, sadness and surprise [6]. Most research related to emotional classification in Computer science research makes the basic assumption of primary emotions.

Chapter 2: Literature Review

The primary question is whether it is possible to find distinct voice profiles for different emotions, such that the voice can be used to infer what the speaker is feeling. This has proved difficult due to both practical problems and the complex nature of the voice production process. There are several sources of variability that complicate the search for voice profiles, such as individual differences among speakers, effects of the verbal content, interactions between spontaneous and strategic expression, as well as important variations within particular emotion families (e.g., hot vs. cold anger). Predictably, reviews of the literature commonly mention inconsistent data regarding voice cues to specific emotions. Only correlates of overall arousal level such as high F0 and fast tempo are very consistently replicated, which has led some to propose that only arousal is coded in voice. However, there is considerable evidence that voice cues can differentiate affective states beyond the simple affective dimensions of activation (aroused/sleepy) and valence (pleasant/unpleasant). Most studies have used emotion portrayals by professional actors and a crucial question is to what degree such portrayals differ from natural expressions. The jury is still out because few attempts have been made to directly compare the two types of speech samples. In addition, most studies have focused on only a few "basic emotions", while neglecting more complex emotions. Hence, much of the pertinent work on emotion differentiation in the voice remains to be done.

In [7], the author provides an up to date record of available emotional speech data collections, the number of emotional states, the language, the number of speakers and the kind of speech. Typical features are the pitch, the formants, the vocal-tract cross sectional areas, Mel-frequency Cepstral coefficients, the Teager energy operator based features, the intensity of speech signal, and the speech rate. Hidden Markov Models, Artificial Neural Networks, Linear discriminant analysis, k nearest neighbours, and support vector machines are used to classify five emotional states i.e. Anger, Disgust, Fear, Joy and Sadness.

By training emotion-specific Language arid prosodic models on a corpus consisting of several thousands of sad, angry, or neutral speech segments from English movies, they showed that a classification system based on these models achieved accuracy comparable to the accuracy of human listeners performing the same task [8].

Emotion modulates almost all modes of human communication word choice, tone of voice, facial expression, gestural behaviours, posture, skin temperature, clamness, respiration, muscle tension, and more. Emotions can significantly change the message: sometimes it is not what was said that was most important, but how it was said. Physiological pattern recognition of emotion has important applications in medicine, entertainment, and human computer interaction. Affective states of depression, anxiety, and chronic anger have been shown to impede the work of the immune system, making people more vulnerable to viral infections, and slowing healing from surgery or disease. Physiological pattern recognition can potentially aid in assessing and quantifying stress, anger, and other emotions that influence health [9].

In [10], the author explores the dynamics of prosodic parameters (refer to local or fine variations in prosodic parameters) with respect to time. The proposed dynamic features of prosody are represented by: (1) Sequence of durations of syllables in the utterance (duration contour), (2) Sequence of fundamental frequency values (pitch contour) and (3) Sequence of frame energy values (energy contour). Emotion specific vocal tract information is mainly represented by spectral features like Mel-frequency Cepstral coefficients (MFCC), Linear Prediction Cepstral coefficients (LPCC) and their derivatives. The parameters like pitch, duration and energy are used as basic prosodic features, and their derivatives extracted from longer speech segments are used to categorize the emotions present in the speech.

This paper [11] proposes epoch parameters extracted from LP (Linear Prediction) residual and zero frequency filtered speech signal for recognising the emotions present in speech. Instant of glottal closure within pitch period of LP residual is known as an ’epoch’. The significant excitation of vocal tract usually takes place at the instant of glottal closure. In this paper the epoch parameters namely strength of epoch, instantaneous frequency, sharpness of epochs, and slope of strength of epochs are used as features for classification of emotions. These features are extracted from the glottal closure region of LP residual. Vocal folds’ vibration, which is quasi periodic in nature, causes a major excitation to the vocal tract system during the speech production mechanism. Epoch is an instant of vocal fold’s closure. In this study an effort has been made to parameterize the epochs, and 4 parameters namely - strength of epochs, instantaneous frequency, sharpness of epochs and slope of strength of epochs are used.

In [12], the author discusses the discriminating capability of a set of features for emotional speech recognition. A total of 87 features were calculated over 500 utterances from the Danish Emotional Speech database. The Sequential Forward Selection method (SFS) was used in order to discover a set of 5 to 10 features which are able to classify the utterances in the best way. The criterion used in SFS is the cross-validated correct classification score of one of the following classifiers: nearest mean and Bayes classifier where class pdfs are approximated via Parzen windows or modelled as Gaussians. After selecting the 5 best features, they reduced the dimensionality to two by applying principal component analysis. The (SFS) algorithm is used for automatic feature selection. The criterion employed was the correct classification rate achieved by the selected features.

To solve the speaker independent emotion recognition problem, the author in [13] proposes a three-level speech emotion recognition model to classify six speech emotions, including sadness, anger, surprise, fear, happiness and disgust from coarse to fine. For each level, appropriate features are selected from 288 candidates by using Fisher rate which is also regarded as input parameter for Support Vector Machine (SVM). There are two key phases to recognize emotions in speech signals, one is to find effective speech emotion features, the other is to establish proper mathematic model for speech emotion recognition. In addition, some high-frequency features can also express some kinds of emotions, such as deep breath sounds for anger and tremble voice for fear.

In [14], Prosodic analysis of speech segments is performed to recognise emotions. Speech signal is segmented into words and syllables. Energy and pitch parameters are extracted from utterances, words and syllables separately to develop emotion recognition models. Eight emotions (anger, disgust, fear, happy, neutral, sad, sarcastic, and surprise) of simulated emotion speech corpus, IITKGPSESC are used in this work for recognition of emotions. Recognition performance of emotions using segmental level prosodic features is not found to be appreciable, but by combining spectral features along with prosodic features, emotion recognition performance is considerably improved.

Since emotional speech can be regarded as a variation on neutral (non-emotional) speech, it is expected that a robust neutral speech model can be useful in contrasting different emotions expressed in speech. This study [15] explores this idea by creating acoustic models trained with spectral features, using the emotionally-neutral TIMIT

corpus. Detecting and utilizing non-lexical or paralinguistic cues from a user is one of the major challenges in the development of usable human-machine interfaces (HMI). Notable among these cues are the universal categorical emotional states (e.g., angry, happy, sad, etc.), prevalent in day-to-day scenarios. Knowing such emotional states can help adjust system responses so that the user of such a system can be more engaged and have a more effective interaction with the system. For the aforementioned purpose, identifying a user's emotion from the speech signal is quite desirable since recording the stream of data and extracting features from this modality is comparatively easier and simpler than in other modalities such as facial expression and body posture. In this report, the TIMIT database was used to train neutral acoustic models and two emotional speech databases were probed. Hidden Markov Models (HMMs) are trained with two different acoustic feature sets, Mel Filter Bank (MFB) and Mel-Frequency Cestrum Coefficients (MFCCs), and their behaviours are examined in a broad phonetic-class recognition experiment setting based on recognition likelihood scores.

Chapter 3: Speech Production Models and Feature Set

3.1: Linear Discrete time Speech Model

In order to apply digital signal processing techniques to speech processing problems it is essential to understand the fundamentals of the speech production process. Speech signals are composed of sequence of sounds. These sounds and the transition between them serve as a symbolic representation of the information [16].

The human vocal tract consists of the pharynx and the mouth or oral cavity. In average male, the total length of vocal tract is 17cm. The cross-sectional area of vocal tract, determined by the positions of the tongue, lips, jaw, and velum varies from zero to about 20. The nasal tract begins at the velum and ends at the nostrils. When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech. The sub-glottal system, consisting of lungs, bronchi and trachea serves as a source of energy for the production of speech. Speech is simply acoustic wave that is radiated from this system when air is expelled from the lungs and the resulting flow of air is perturbed by a constriction somewhere in the vocal tract. Speech signals can be classified in three distinct classes according to mode of excitation.

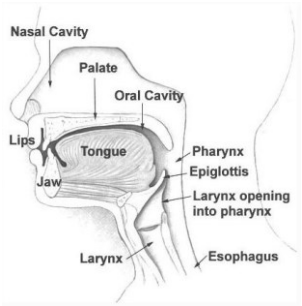


Fig 3.1.1: *Vocal tract*

• **Voiced sounds:** are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of air which excite the vocal tract.

• **Unvoiced sounds:** are generated by forming a constriction at some point in the vocal tract and forcing air through the constriction at a high enough velocity to produce turbulence. This creates a broad spectrum noise source to excite the vocal tract.

• **Plosive sounds:** result from making a complete closure, building up pressure behind the closure and abruptly releasing it.

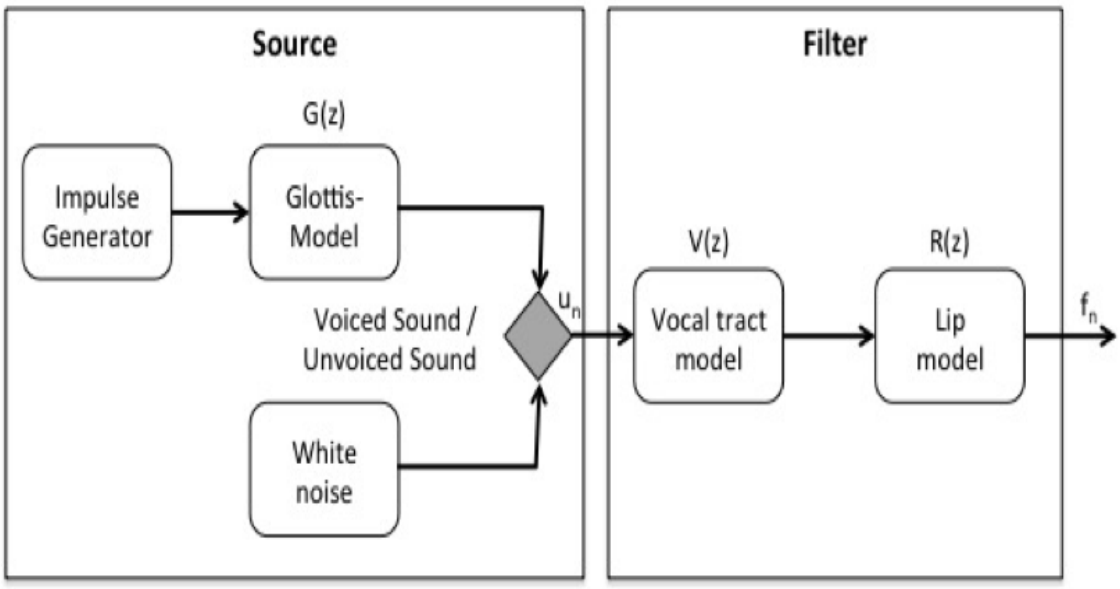


Fig 3.1.2: *General discrete time linear model for speech production*

3.1.1: Linear Speech Model based Feature Set: The below mentioned features are extracted based on the assumption of linear speech production model.

- **Short time Energy [16]:** The amplitude of unvoiced segments is generally lower than the amplitude of voiced segments. The short time energy of a speech signal provides a convenient representation that reflects these amplitudes variations. The short time energy is defined as

$$E_n = \sum_{m=-\infty}^{\infty}[x(m)w(n-m)]^2 \qquad \textit{Eq. 3.1}$$

$$where\ x(m) is\ speech\ sample\ and\ w(n) is\ a\ window\ function$$

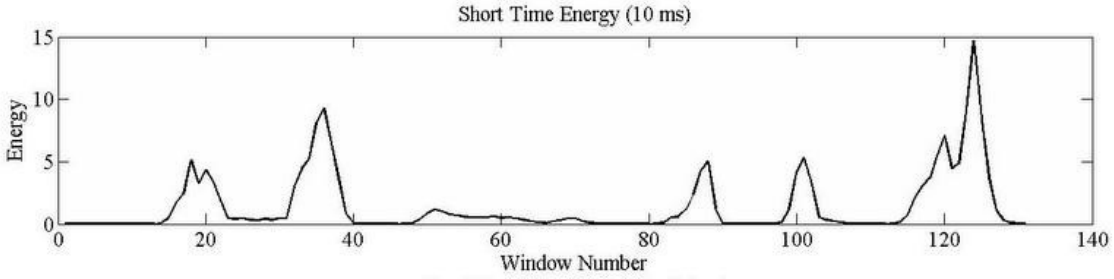


Fig 3.1.3: *Short time energy profile for a speech sample*

- **Short time average magnitude [16]:** A problem with short time energy function is that it is very sensitive to large signal values which emphasizes a lot on sample to sample variation of x(n). Short time Average Magnitude function elevates this problem as

$$M_n = \sum_{m=-\infty}^{\infty} |x(n)|w(n-m) \qquad \textbf{Eq. 3.2}$$

where x(n) is speech sample and w(n) is a window function.

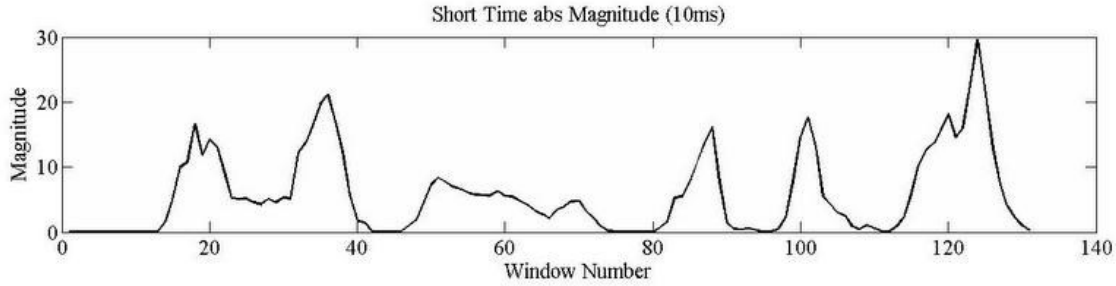


Fig 3.1.4: *Short time absolute magnitude profile for a speech sample*

- **Short time average zero cross rate [16]:** A zero crossing defines the number of times the sign of signal changes within a specified window. The rate at which zero crossing occurs is simply a measure of frequency content of the signal. This is particularly true for narrowband signals with zero dc bias. Speech signals are broadband signals and hence interpretation of zero-crossing rate is therefore much less precise. However, rough estimates of spectral properties can be obtained for it. The definition of Zero Crossing rate is

$$Z_n = \sum_{m=-\infty}^{\infty} \left|sgn\big(x(m)\big) - sgn\big(x(m-1)\big)\right| w(n-m) \qquad \textbf{Eq. 3.3}$$

where $sgn\big(x(n)\big) = 1, if x(n) > 0\,, else - 1$

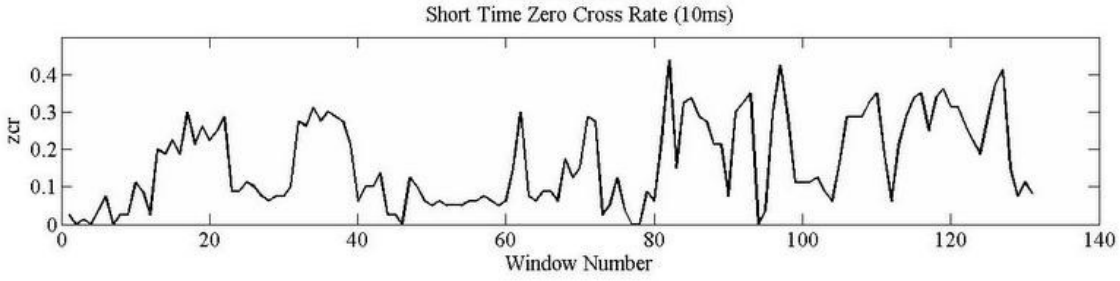


Fig 3.1.5: *Short time zero cross rate for a speech sample*

- **Pitch estimation (Using autocorrelation function) [17],[18],[19]:** Autocorrelation of a speech signal has many peaks and most of these peaks can be attributed to the damped oscillations of the vocal tract response which are responsible for the shape of each period of speech wave. Formant frequencies can also cause a problem as their peak might be higher than the pitch period peak. Hence, to avoid above problems we need some pre-processing of the speech signals. This process is known as spectrum flatteners. The algorithm to detect pitch of a speech using autocorrelation method is based on clipping.

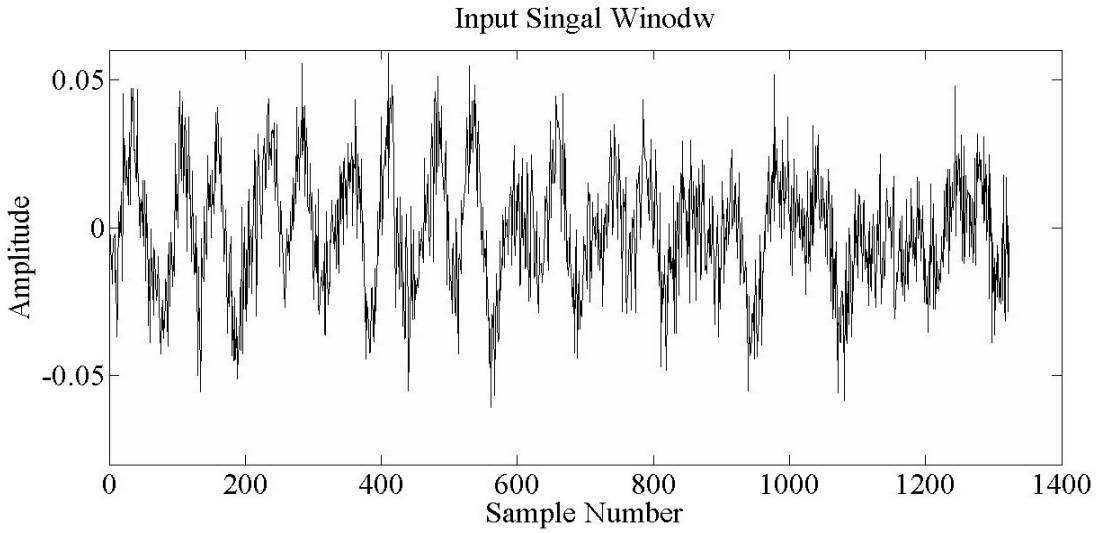


Fig 3.1.6: *Input Speech Sample*

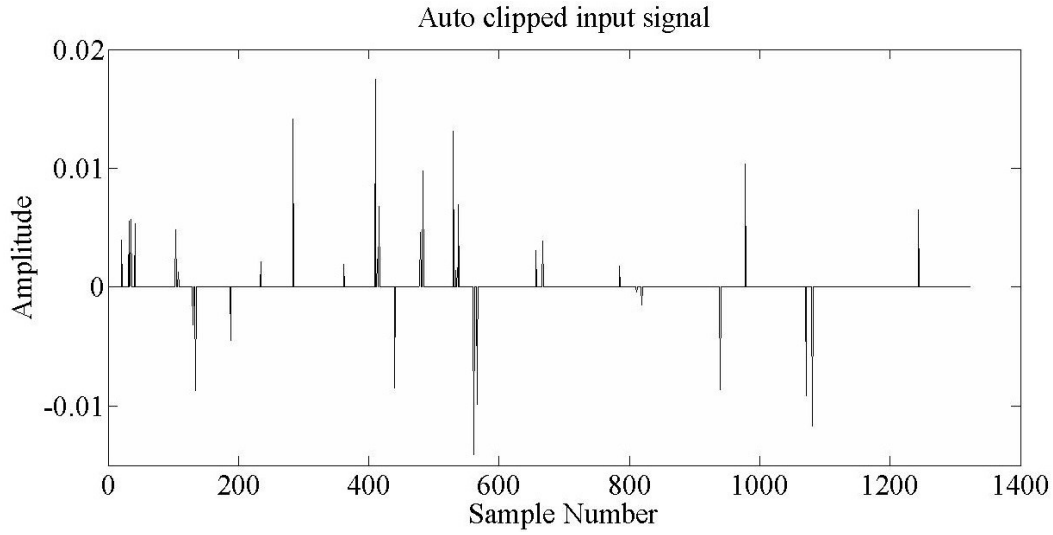


Fig 3.1.7: *Auto clipped input signal*

Steps:

1. Break the signal into windows of 25ms with an overlap of 10 ms.
2. For each window calculate the maximum and minimum amplitude point.
3. Clip the signal on positive side at 0.8 times of maximum value and on negative side on 0.8 times of minimum value.
4. Find autocorrelation of the clipped signal.
5. Find the peak in the autocorrelation function, other than the one present with zero lag. The point amounts to pitch of signal. Convert it to frequency (in Hz).

- **Linear Predicative Coding of Speech (LPC) [16]:** The basic idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. By minimizing the sum of the squared differences between the actual speech samples and the linearly predicted ones, a unique set of predictor coefficients can be determined. The various algorithms that are used for LPC calculations are
    1. Covariance Method
    2. Autocorrelation Method
    3. Inner product Method

  The composite effect of radiation, vocal tract, and glottal excitation are represented by a time varying digital filter whose steady state function s of the form

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-1}} \qquad \boldsymbol{Eq.3.4}$$

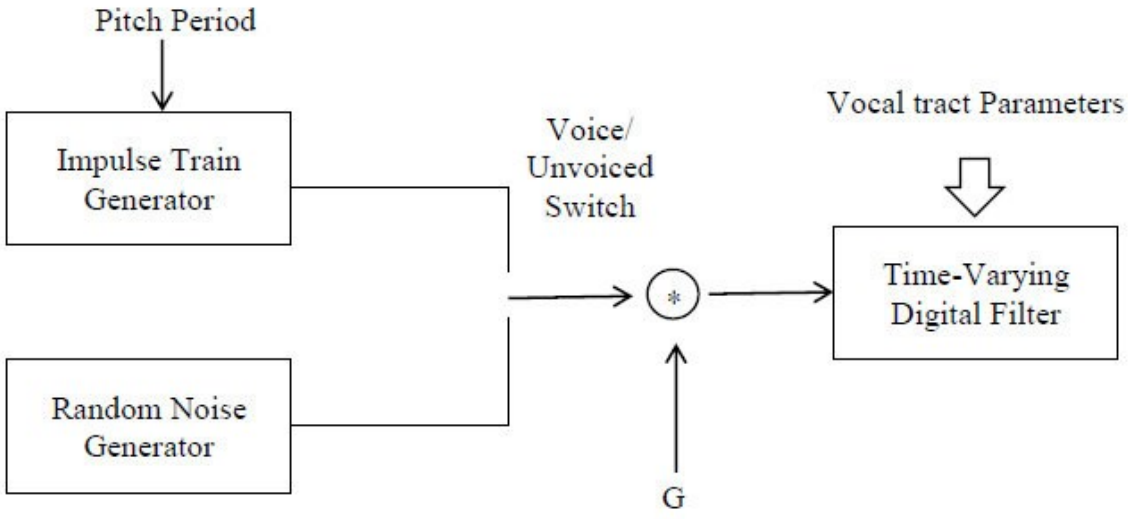


Fig 3.1.8: *Block Diagram of simplified Speech Production Model*

- **Formant Frequencies:** Formants are defined by Gunnar Fant as "the spectral peaks of the sound spectrum |P(f)|". In acoustics generally, a very similar definition is widely used: the Acoustical Society of America defines a formant as: "a range of frequencies [of a complex sound] in which there is an absolute or relative maximum in the sound spectrum. In speech science and phonetics, however, a formant is also sometimes used to mean an acoustic resonance of the human vocal tract. Thus, in phonetics, formant can mean either a resonance or the spectral maximum that the resonance produces. Formants are often measured as amplitude peaks in the frequency spectrum of the sound, using a spectrogram or a spectrum analyser and, in the case of the voice; this gives an estimate of the vocal tract resonances. In vowels spoken with a high fundamental frequency, as in a female or child voice, however, the frequency of the resonance may lie between the widely-spaced harmonics and hence no corresponding peak is visible.
    - Formants are considered to be robust against channel distortions and noise.
    - Formant parameters might provide a means to tackle the problem of a mismatch between training and testing conditions.
    - There is a close relation of formant parameters to model based approaches to speech perception and production.

3.2: Non - Linear Discrete time Speech Model [20], [27], [28], [29]

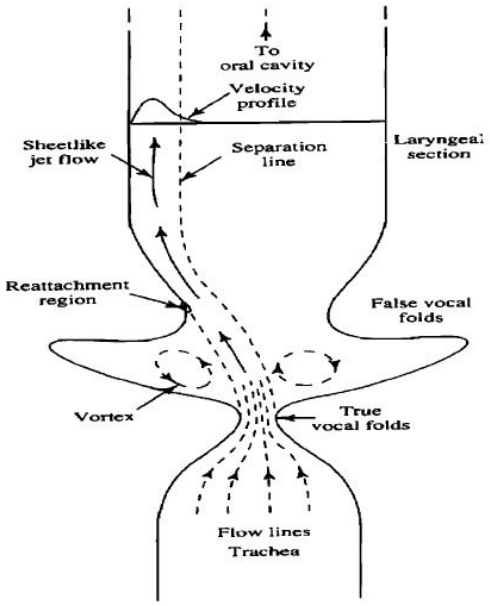


Fig 3.2.1: *Nonlinear model of sound propagation along the vocal tract*

In the speech production process, there is a net airflow through the glottis. The linear acoustic model of speech production says that this flow only causes sound when forced through a constriction (i.e., fricative production). However, if the propagation of the glottal flow through the vocal tract created vortices of air in the region of the false vocal folds, sound could be actively produced from a source other than the glottis.

This phenomenon of sound creation by vortex action is nonlinear and cannot be measured by any of the techniques employed to date. Teager [27], [28], [29], who suggested that these vortices modulated airflow in the vocal tract causing sound, developed the Teager Energy operator. The operator was used to show modulation patterns in the energy of individual formants vortices modulated airflow in the vocal tract causing sound, developed the Teager Energy operator. The operator was used to show modulation patterns in the energy of individual formants. In this model, air exits the glottis as a jet and attaches to the nearest wall of the vocal tract. As the air passes over the cavity between the true vocal folds and the false vocal folds, vortices of air are created. The bulk of the air continues propagating towards the lips while adhering to the walls of the vocal tract. The key element in this model is the vortex action. A traditional model of speech production allows sound to be actively produced in an un-constricted vocal tract only at the glottis. Teager asserted that vortices in the region of the false vocal folds are also actively producing sound that causes modulations in the speech signal. The continuous from of the TEO is

$$\varphi_c[x(T)] = \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \qquad \boldsymbol{Eq.3.5}$$

$$= \left[\dot{x(t)}\right]^2 - x(t)\ddot{x}(t) \qquad \textbf{Eq. 3.6}$$

Since speech is represented in discrete form in most current speech processing systems, Kaiser [21], [22] derived the operator for discrete-time signals from its continuous form $\varphi[x(n)]$ as:

$$\varphi[x(n)] = x^2(n) - x(n+1)x(n-1) \qquad \textbf{Eq. 3.6}$$

where *x(n)* is the sampled speech version. The TEO is typically applied to a band pass filtered speech signal, since its intent is to reflect the energy of the nonlinear flow within the vocal tract for a single resonant frequency. Although the output of a band pass filter still contains more than one frequency component, it can be considered as an AM–FM signal,

$$r(t) = a(t)\cos(2\pi f(t)t) \qquad \textbf{Eq. 3.7}$$

The TEO output of *r(t)* can be approximated as

$$\varphi[r(t)] = [a(t)2\pi f(t)]^2 \qquad \textbf{Eq. 3.8}$$

In fact, the TEO profile can be used to decompose an AM–FM signal into its AM and FM components within a certain frequency band via

$$f(n) = \frac{1}{2\pi T} arccos\left(1 - \frac{\varphi[y(n)] + \varphi[y(n+1)]}{4\varphi[x(n)]}\right) \qquad \textbf{Eq. 3.8}$$

$$|a(n)| = \sqrt{\frac{\varphi[x(n)]}{\left[1 - \left(1 - \frac{\varphi[y(n)] + \varphi[y(n+1)]}{4\varphi[x(n)]}\right)^2\right]}} \qquad \textbf{Eq. 3.9}$$

Where:

$y(n) = x(n) - x(n-1)$ **Eq. 3.10** ;time domain difference equation.

$\varphi[.]$ = TEO operator.

$f(n)$ = FM Component of signal at n

$a(n)$ = AM Component of signal at n.

On the basis of this work, Maragos et al. proposed a nonlinear model which represents the speech signal as [23]

$$s(t) = \sum_{m=1}^{M} r_m(t) \qquad \textbf{Eq. 3.11}$$

$$r_m(t) = a_m(t)cos\left(2\pi\left(f_{cm}t + \int_0^t q_m(\tau)d\tau\right) + \theta\right) \qquad \textbf{Eq. 3.12}$$

is a combined AM and FM structure representing a speech resonance at the m^{th} formant with a center frequency $F_m = f_{cm}$. In this relation, $a_m(t)$ is the time-varying amplitude, and $q_m(\tau)$ is the frequency modulating signal at the m^{th} formant. Although TEO processing is intended to be used for a signal with a single resonant frequency, but the TEO energy of a multi-frequency signal does not only reflects individual frequency components but also reflects interactions between them. This characteristic extends the use of TEO to speech signals filtered with wide bandwidth band-pass filters (BPF).

- **TEO-FM-Var (Variation of FM Component) [23]:** Voiced speech spoken under stress generally has different instantaneous excitation variations from voiced speech spoken under neutral conditions. This can be verified by comparing voiced speech waveforms spoken under neutral and simulated angry conditions.

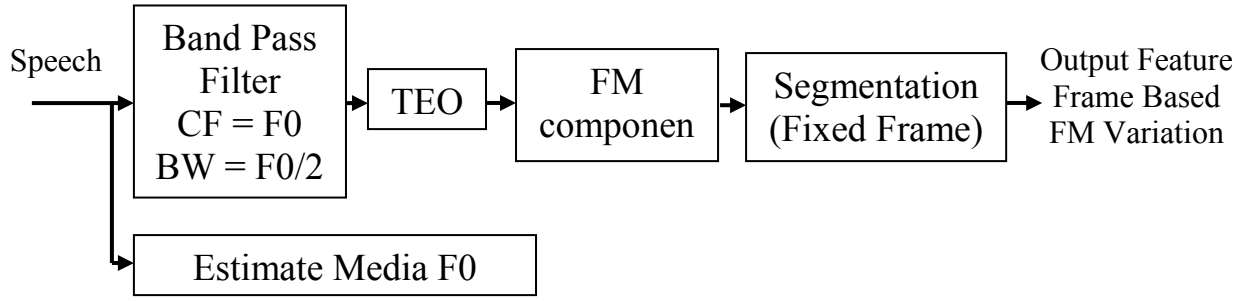


Fig 3.2.2: *TEO-FM Var feature calculation flow*

- **TEO-Auto-Env (Normalized TEO Autocorrelation Envelope Area) [23]:** The motivation for the TEO-FM-Var feature is to capture stress dependent information that may be present in changes within the FM component. Its processing is based on the entire band although the final FM variations are computed around the restricted frequency band. However, the presence of stress may affect modulation patterns across the entire speech frequency band. If a filter bank is used to band-pass filter voiced speech around each of its formant frequencies, the modulation pattern around each formant can be obtained using TEO AM–FM decomposition, from which variations of modulation patterns across different frequency bands can be obtained.

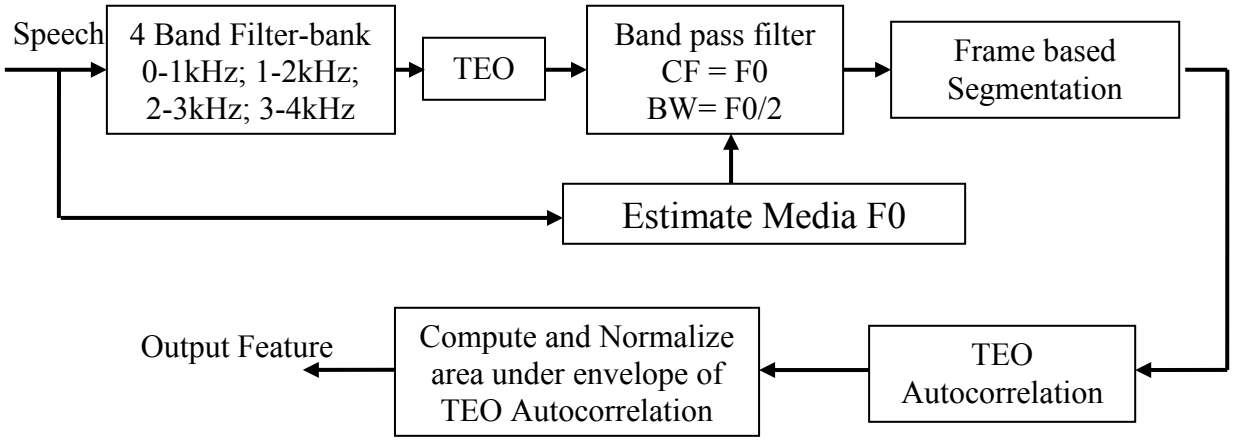


Fig 3.2.3: *TEO autocorrelation Envelope feature extraction flow*

- **TEO-CB-Auto-Env (Critical Band Based TEO Autocorrelation Envelope) [23]:** The uniform partition of the entire speech frequency band for the TEO-Auto-Env was performed in an attempt to capture stress sensitive changes outside the first formant. The TEOAuto- Env feature allows us to probe nonlinear energy changes at higher frequencies.

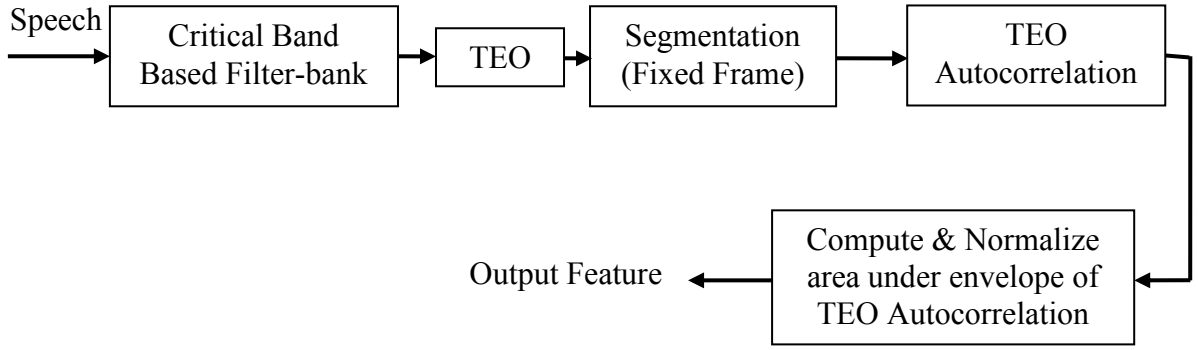


Fig 3.2.4: *TEO critical Bank Autocorrelation Envelope Calculation*

Band Number	Lower f (in Hz)	Central f(in Hz)	Upper f(in Hz)	Band Number	Lower f (in Hz)	Central f(in Hz)	Upper f(in Hz)
1	100	150	200	9	1080	1170	1270
2	200	250	300	10	1270	1370	1480
3	300	350	400	11	1480	1600	1720
4	400	450	510	12	1720	1850	2000
5	510	570	630	13	2000	2150	2320
6	630	700	770	14	2320	2500	2700
7	770	840	920	15	2700	2900	3150
8	920	1000	1080	16	3150	3400	3700

Table 3.2.1: *Critical Bank filter specifications*

Chapter 4: Results

4.1: Dataset Description [30] (German Emotional Database): Ten actors (5 female and 5 male) simulated the emotions, producing 10 German utterances (5 short and 5 longer sentences) which could be used in every day communication and are interpretable in all applied emotions. The recordings were taken in an anechoic chamber with high-quality recording equipment. The speech material comprises about 800 sentences (seven emotions * ten actors * ten sentences + some second versions). The complete database was evaluated in a perception test regarding the recognisability of emotions and their naturalness. Utterances recognised better than 80% and judged as natural by more than 60% of the listeners were phonetically labelled in a narrow transcription with special markers for voice-quality, phonatory and articulatory settings and articulatory features.

4.1.1: Information about speakers: {N = Neutral, A = Anger, B = Boredom, D = Disgust, F = Fear, H = Happiness, S = Sadness}

Speaker Code	Sex	Age	No of Emotions samples [N,A,B,D,F,H,S]
03	Male	31	(11,14,5,1,4,7,7)
08	Female	34	(10,12,10,0,6,11,9)
09	Female	21	(9,13,4,8,1,4,4)
10	Male	32	(4,10,8,1,8,4,3)
11	Male	26	(9,11,8,2,10,8,7)
12	Male	30	(4,12,5,2,6,1,4)
13	Female	32	(9,12,10,8,7,10,5)
14	Female	35	(7,16,8,8,12,8,10)
15	Male	25	(11,13,9,5,8,7,4)
16	Female	31	(5,14,14,11,7,11,9)

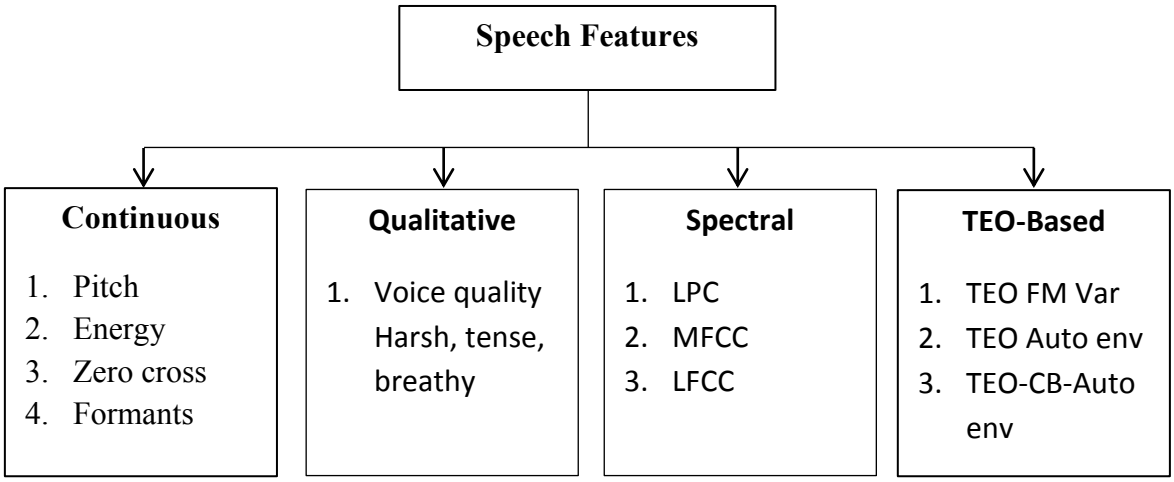
Table 4.1.1: *Speaker information about Berlin Emotional Database*

4.1.2: Text material: Setting up the database with speech data in which actors simulate emotions has the advantage that it is possible to control the individual sentences to be spoken. It is important, though, that all these sentences should be interpretable in the emotions under review and that they contain no emotional bias. Two different kinds of text material would normally meet these requirements:

- Nonsense text material, like for instance haphazard series of figures or letters, or fantasy words.
- Normal sentences which could be used in everyday life.

Nonsense material is guaranteed to be emotionally neutral. Recordings were taken with a sampling frequency of 48 kHz and later down sampled to 16 kHz. The actors were standing in front of the microphone so they could use body language if desired; only hindered by the cable of the laryngograph and the need to speak in the direction of the microphone with a distance of about 30 cms.

4.2: Feature set:



4.3: Discrimination between Speech and Silence: To extract the feature set, the silence part in the recorded waveform must not be processed. Hence we used a speech-silence classifier based on Short Time Energy and Short Time Zero Cross Rate. The problem of locating the beginning and end of a speech utterance in a background noise is of importance in many areas of speech processing. A scheme for locating the beginning and end of a speech signal can be used to eliminate significant computation in real-time systems by making it possible to process only the parts of the input that correspond to speech. The algorithm given below works on the principle that first 100msec of speech data is pure noise. Even if the above case is not true then for most of real time applications we have a separate mike to record the noise profile after a fixed interval and that can be utilized effectively with no change in algorithm. The algorithm can be quite useful when high rate of processing is required for ex. in speech recognition, as the calculations can be reduced by at least one thirds for normal speaker. The other use of this algorithm is noise

removal. Most noise removal algorithms require a sample of noise after a fixed interval to update their references for noise removal.

Algorithm:

Step 1: Calculate mean zero cross rate and Short time energy for a noise profile with a window length of 10 ms.

Step 2: For a signal window of 10 ms , if the energy is greater than C times energy of noise signal and zero cross rate is also greater than zero cross rate for noise signal, then it is a speech window, else noise.

Step 3: Update noise zero cross rate and energy every 5 to 10 seconds, to take varying noise in consideration.

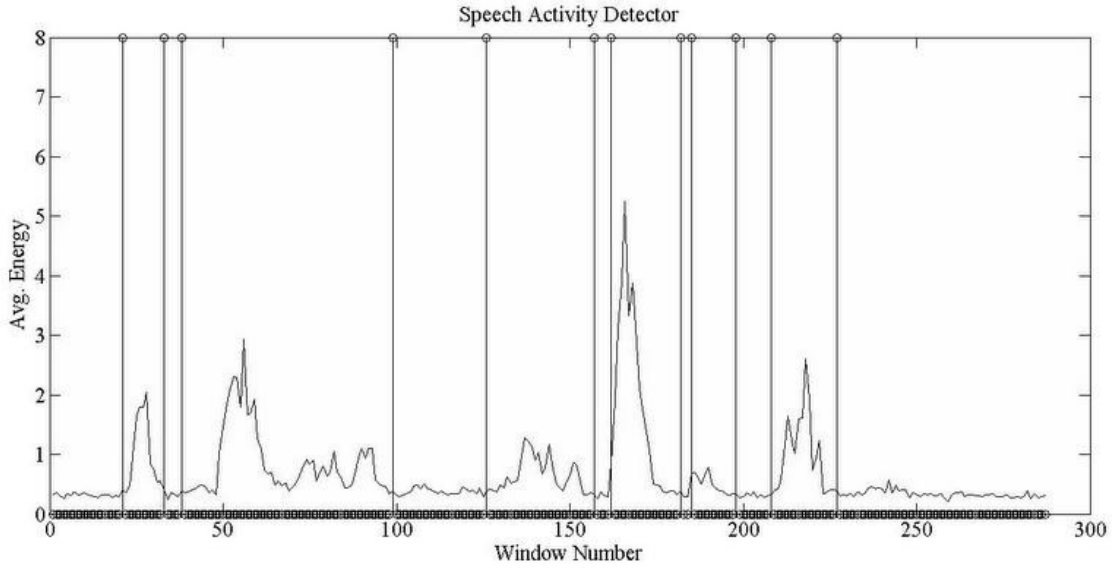


Fig 4.3.1: *Speech-Silence classifier*

	Pitch				Intensity	
	Mean	Range	Variance	Contour		
Anger	>>	>	>>		>>	>
Disgust	<	>m, <f			<	
Fear	>>	>		↗	>=	
Joy	>	>	>	↘	>	>
Sadness	<	<	<	↗	<	<

Table 4.3.1: *Observed Properties for different emotions.*

4.4: TEO decomposition of various Emotions for word: abgeben

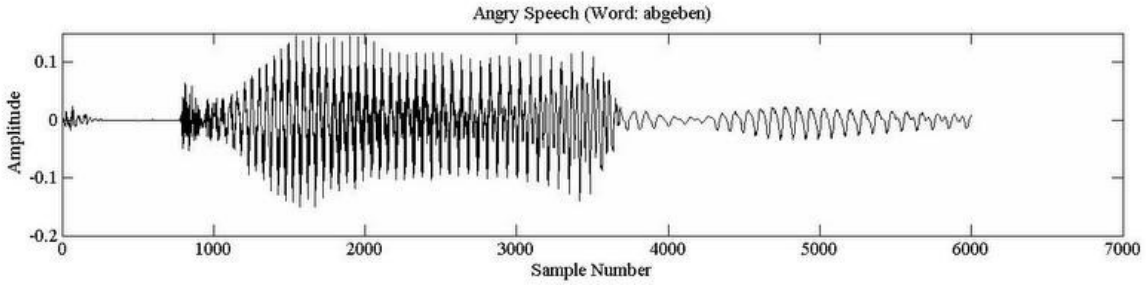


Fig 4.4.1: *TEO Decomposition of Angry Emotion*

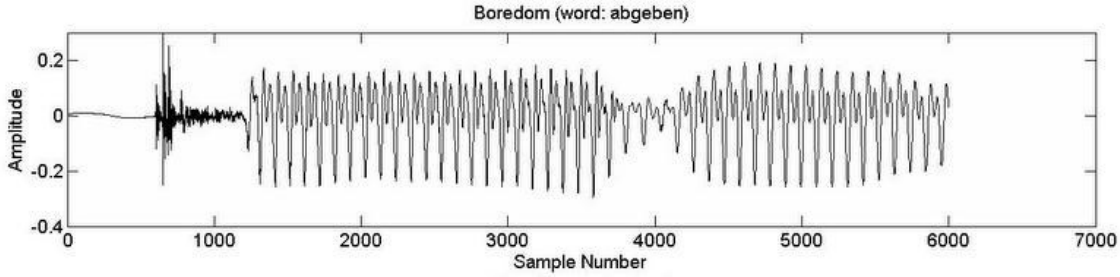


Fig 4.4.2: *TEO Decomposition of Boredom Emotion*

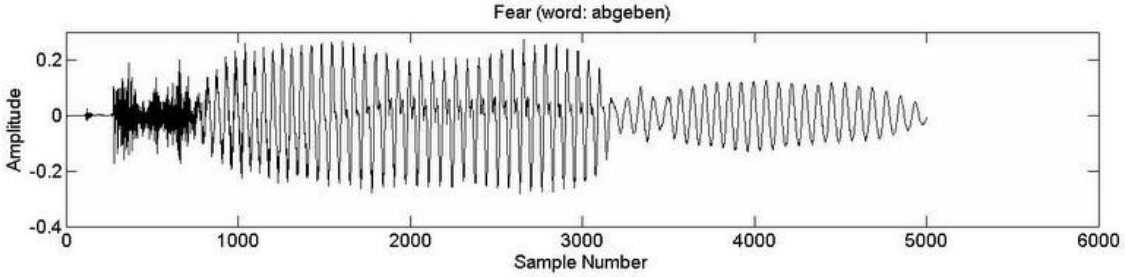


Fig 4.4.3: *TEO Decomposition of Fear Emotion*

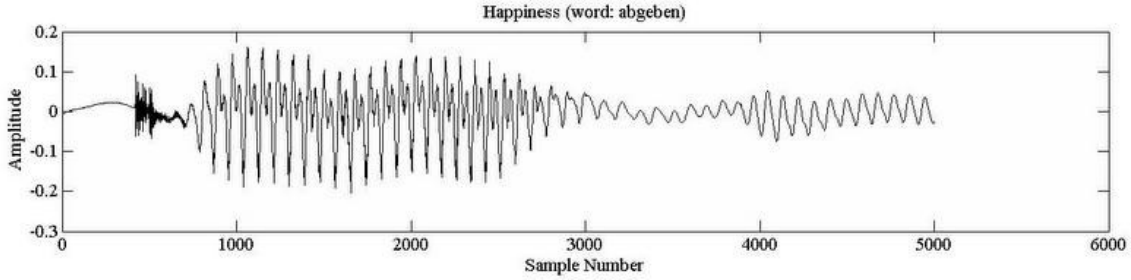


Fig 4.4.4: *TEO Decomposition of Happiness Emotion*

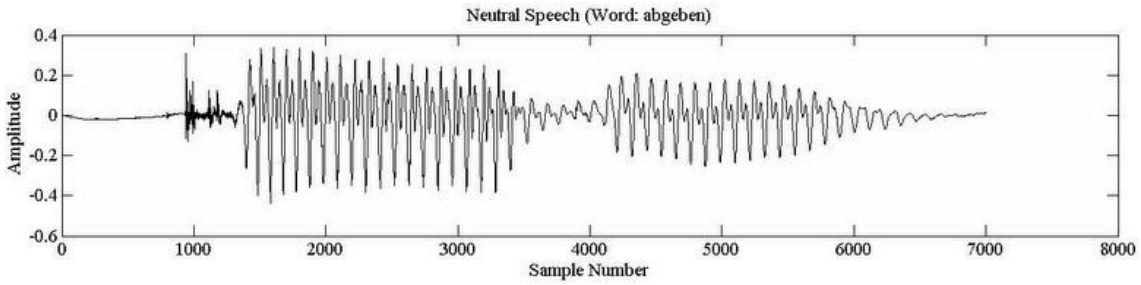


Fig 4.4.5: *TEO Decomposition of Neutral Emotion*

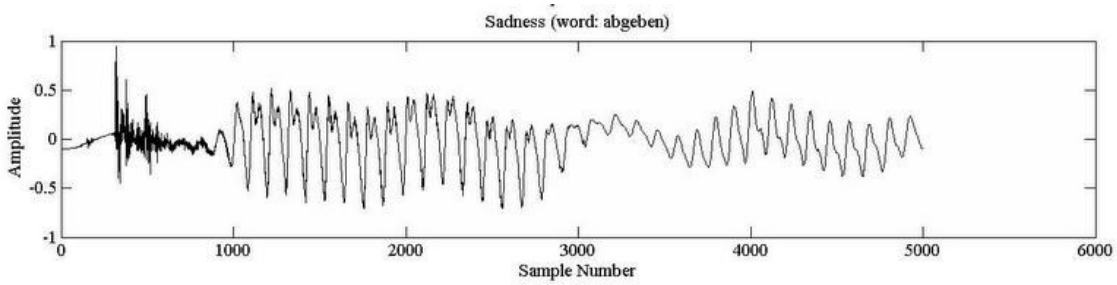


Fig 4.4.6: *TEO Decomposition of Sadness Emotion*

4.5: FM (Frquency Modulation) decomposition of various Emotions for word: abgeben

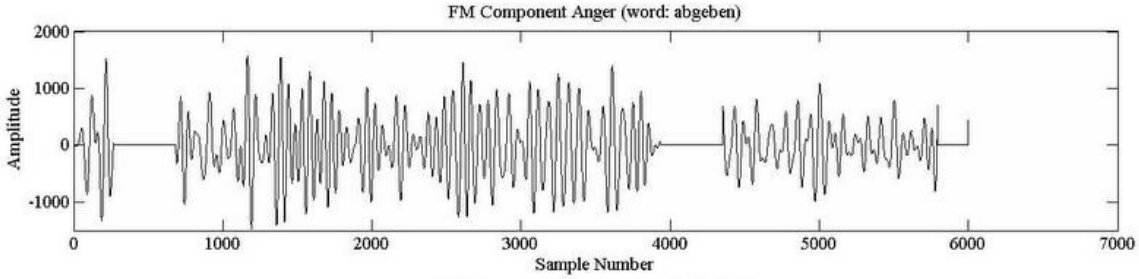


Fig 4.5.1: *FM Decomposition of Anger Emotion*

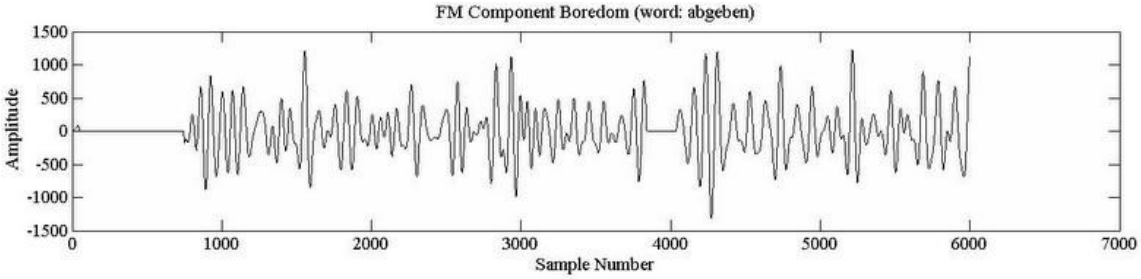


Fig 4.5.2: *FM Decomposition of Boredom Emotion*

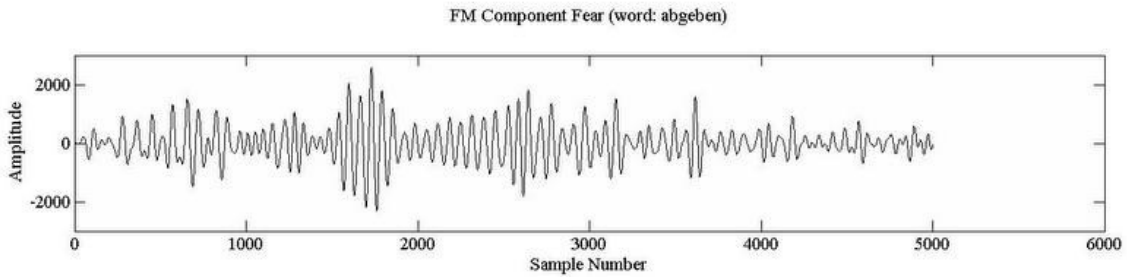


Fig 4.5.3: *FM Decomposition of Fear Emotion*

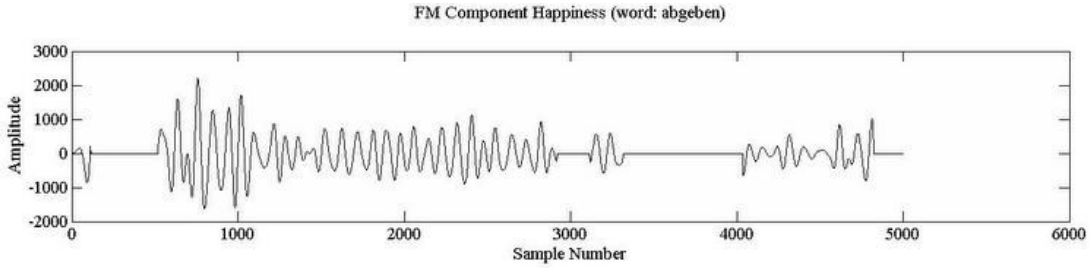


Fig 4.5.4: *FM Decomposition of Happiness Emotion*

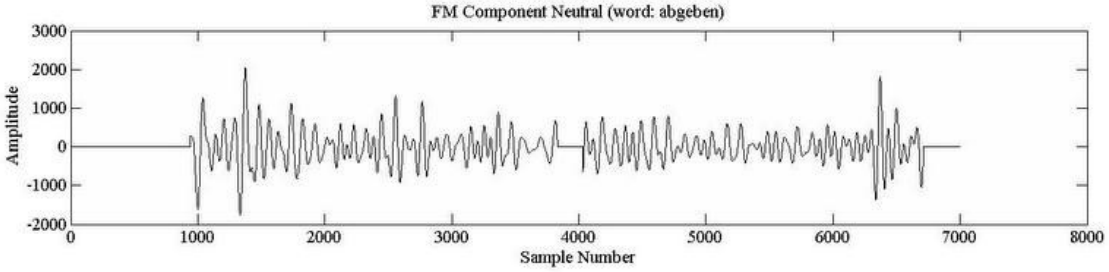


Fig 4.5.5: *FM Decomposition of Neutral Emotion*

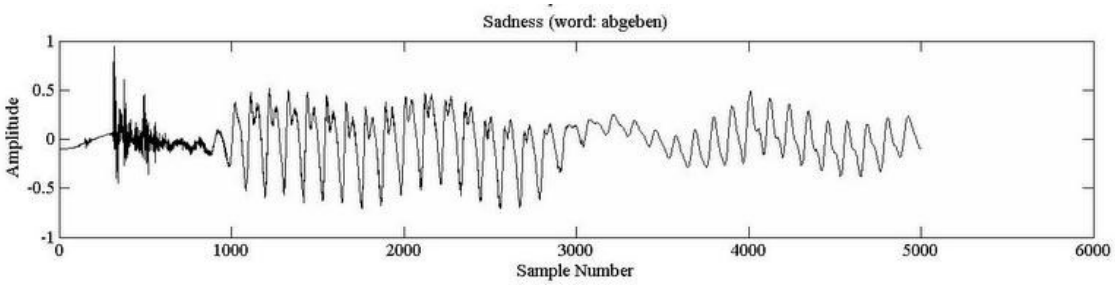


Fig 4.5.6: *FM Decomposition of Sadness Emotion*

4.6: Results: ANN-based classifiers are used for emotion classification due to their ability to find non-linear boundaries separating the emotional states. The most frequently used class of neural networks is that of feed-forward ANNs, in which the input feature values propagate through the network in a forward direction on a layer-by-layer basis. Typically, the network consists of a set of sensory units that constitute the input layer, one or more hidden layers of computation nodes, and an output layer of computational nodes. Let us consider a one-hidden layer feed-forward neural network that has Q input nodes, A hidden nodes, and B output nodes. An interesting property of ANNs is that by changing the number of hidden nodes and hidden layers we control the non-linear decision boundaries between the emotional states.

The neural network provides a mapping of the form $z = f(y)$ defined by;

$$v_a = g_1(w_a^T y + w_0) \qquad \boldsymbol{Eq.4.1}$$

$$z_b = g_2(u_b^T v + u_0) \qquad \boldsymbol{Eq.4.2}$$

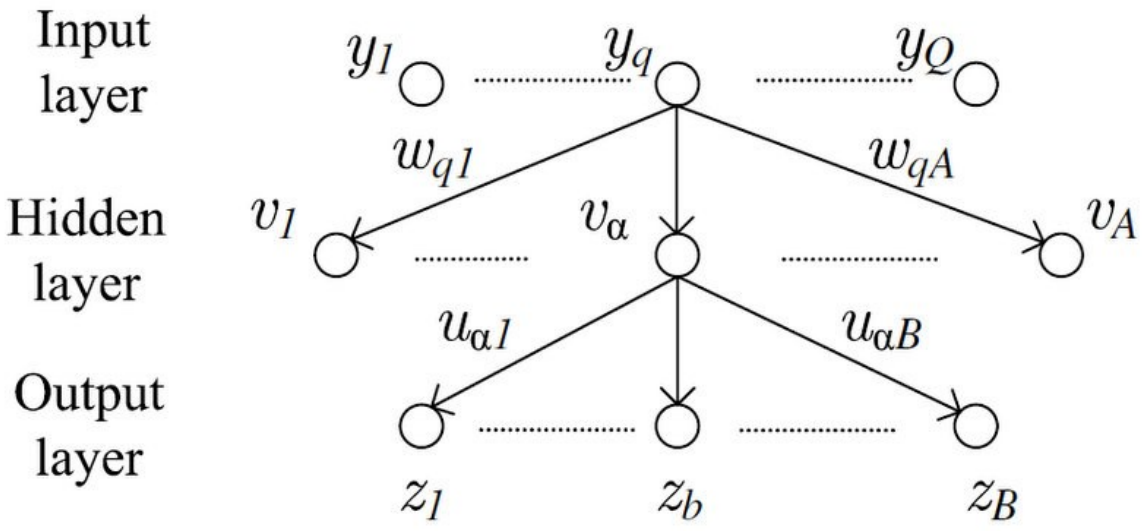


Fig 4.6.1: *Feed-forward Neural Network with one hidden layer*

In first case, the number of output nodes of ANN was equal to number of emotional states (i.e. seven). The feature set included two classes, Linear and Non-Linear. In first case, the

ANN was trained using combined feature set (both linear and non-linear). The accuracies obtained was in range of 60 to 70% for a dataset containing six emotions. In second case, only linear feature set was used and classification accuracy remained within 60 to 70%. The standard deviation observed in accuracies of combined feature set was lower than just linear feature set. Also, Disgust and Anxiety showed lowest classification accuracy when using just linear feature set.

Combined (Linear and Non-Linear features)								
% data	Neurons	Overall	Anger	Anxiety	Boredom	Disgust	Happiness	Sadness
50	30	64.15	72.09	22.58	39.47	56.52	63.64	89.58
60	30	68.37	77.50	52.38	78.26	24.00	78.57	90.00
70	30	70.00	68.00	54.55	57.14	80.00	56.52	78.57
80	30	69.81	64.71	38.46	61.54	50.00	57.14	100.00
90	30	61.11	36.36	66.67	85.71	60.00	37.50	100.00
Linear								
50	60	62.64	72.09	29.03	47.37	21.74	48.48	95.83
60	60	71.16	85.00	42.86	69.57	28.00	67.86	85.00
70	80	63.12	64.00	40.91	66.67	30.00	56.52	82.14
80	60	63.20	47.06	38.46	53.85	20.00	50.00	93.75
90	60	61.11	81.82	33.33	57.14	80.00	37.50	77.78

Table 4.6.1: *Accuracies for single ANN Structure.*

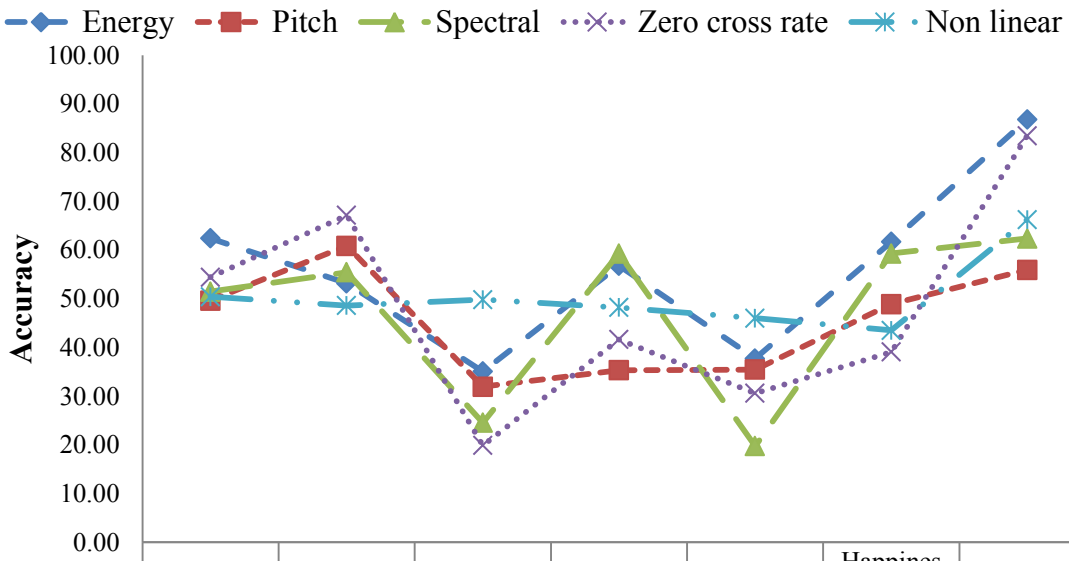
males	females	sb 03	sb 08	sb 09	sb 10	sb 11	sb 12	sb 13	sb 14	sb 15	sb 16
0.63	0.68	0.63	0.76	0.64	0.67	0.70	0.56	0.74	0.73	0.59	0.57

Table 4.6.2: *Accuracies for various speakers*

Disgust shows the least accuracy in case of linear feature set, but when non-linear feature set is combined, the accuracy improves by 20%. In this dataset, the female were better able to produce the emotions, compared to male speakers and there is difference of 5% accuracy in emotion classification. Also some speakers like sb12 showed an accuracy of just 56%, whereas others like sb08 showed accuracy of 76%. The dataset was selected based on the speaker emotion recognition by humans of more than 60%.

4.6.1: Different Feature Set Based Classification: Different features could be used to differentiate between different emotions. In this case, we studied the accuracy of emotion on different feature set size. The considered feature sets was:

- Short time Energy
- Pitch
- Spectral Features (Spectral flux, Spectral Roll Off, Spectral Centroid)
- Short Time Zero Cross Rate
- Non Linear Feature set.



	Overall	Anger	Anxiety	Boredom	Disgust	Happiness	Sadness
Energy	62.40	53.14	35.03	56.79	37.71	61.67	86.84
Pitch	49.59	60.86	31.91	35.31	35.43	48.89	55.92
Spectral	51.45	55.43	24.52	59.27	19.71	59.27	62.35
Zero cross rate	54.39	67.14	19.85	41.61	30.57	39.04	83.47
Non linear	50.40	48.57	49.81	48.18	46.00	43.48	66.22

Fig 4.6.2: *Accuracy of different emotions w.r.t to different feature set.*

Just Short time Energy based feature set offers highest accuracy of 62.40%. Though non-linear offers an accuracy of just 50.40%, with least standard deviation 8.07%. All the above are weak classifiers, and could be used in Ensemble classification to increase the classification accuracy. The feature vector size is 40,10,30,40 and 56 in case of Energy, Pitch, Spectral, Zero Cross Rate and Non-Linear respectively. It can be concluded that each emotion affects the linear as well as Non-Linear speech production Model differently.

4.6.2: Group Division Classification: Based on Ekman's six kinds of basic emotion model [31] and Fox's multi-level emotional classification model [13], we have established a three level emotion recognition model which contains five classifiers for pairwise emotion classification. The model have three levels, the first level contains one classifier which divides utterance into two classes. The next two levels contain two classifiers respectively, which further refine utterance.

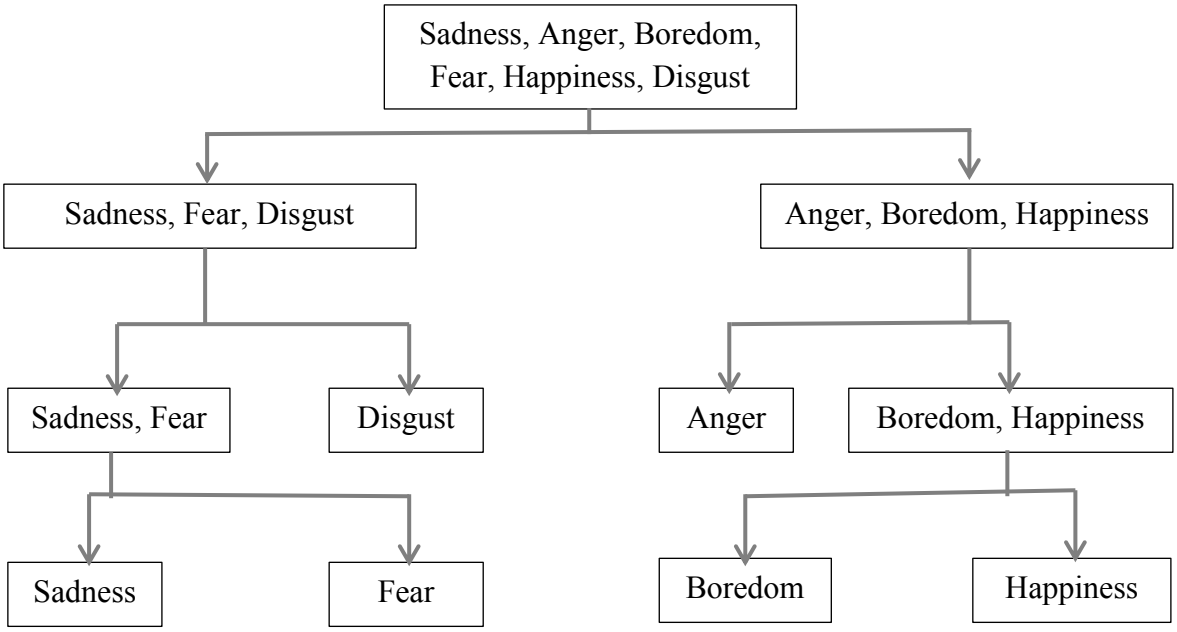
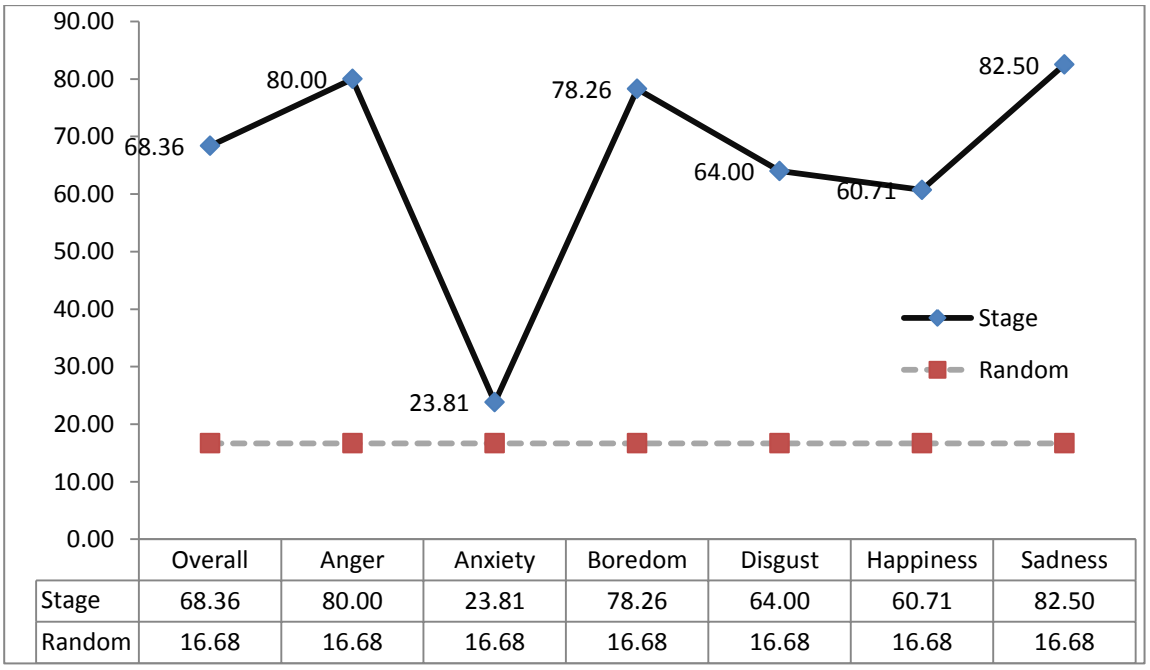


Fig 4.6.3: *Multistage emotion classification approach*



	Overall	Anger	Anxiety	Boredom	Disgust	Happiness	Sadness
Stage	68.36	80.00	23.81	78.26	64.00	60.71	82.50
Random	16.68	16.68	16.68	16.68	16.68	16.68	16.68

Fig 4.6.4: *Group Classification Accuracies*.

4.6.3: Ensemble of Classifiers: Ensemble of classifiers is the concept of combining classifiers for the improvement of the performance of individual classifiers. These classifiers could be based on a variety of classification methodologies, and could achieve different rate of correctly classified individuals. The goal of classification result integration algorithms is to generate more certain, precise and accurate system results. Dietterich [24] provides an accessible and informal reasoning, from statistical, computational and representational viewpoints, of why ensembles can improve results.

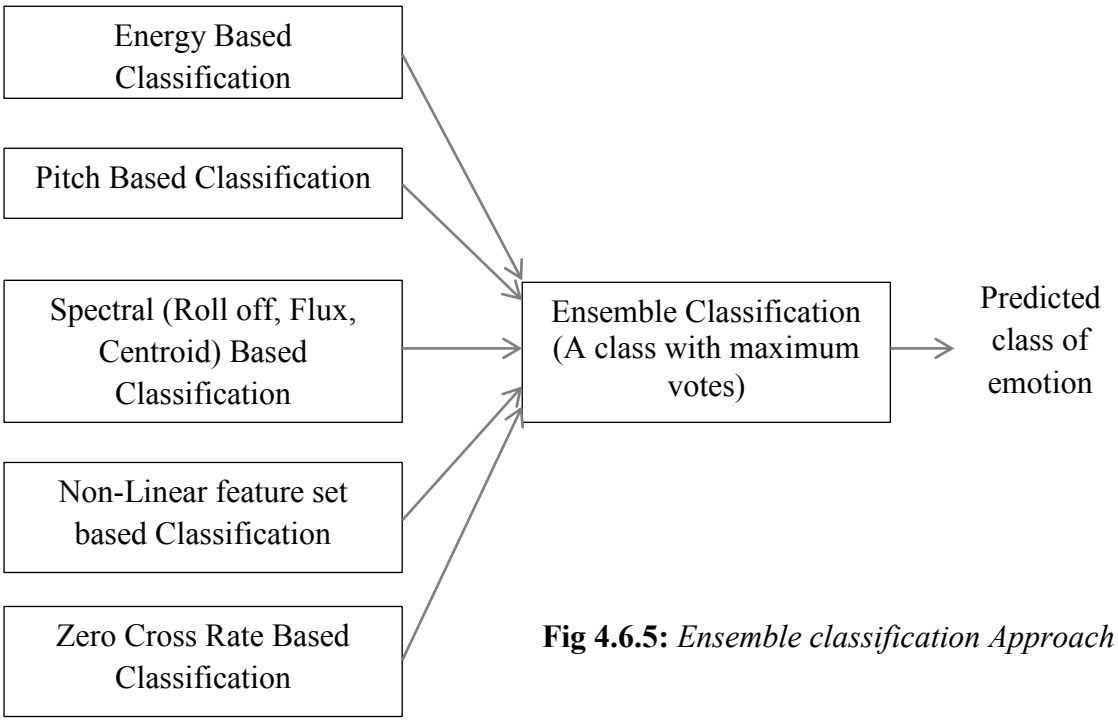
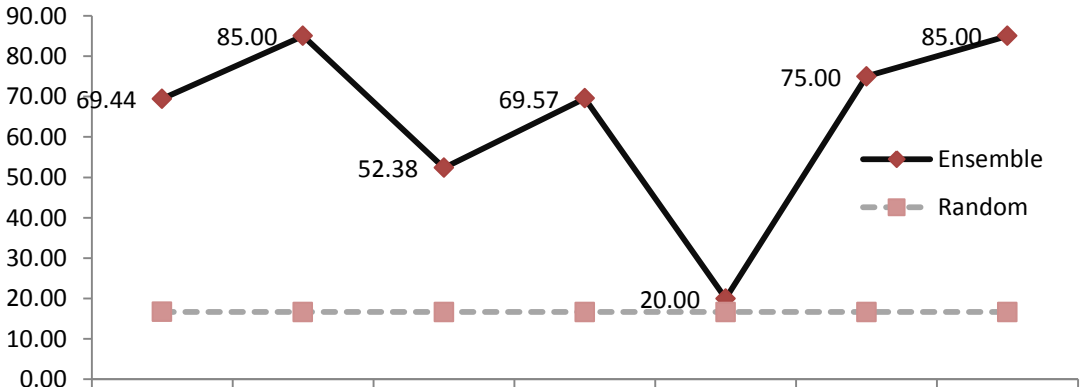


Fig 4.6.5: *Ensemble classification Approach*



	Overall	Anger	Anxiety	Boredom	Disgust	Happiness	Sadness
Ensemble	69.44	85.00	52.38	69.57	20.00	75.00	85.00
Random	16.68	16.67	16.67	16.67	16.67	16.67	16.67

Fig 4.6.6: *Ensemble Classification Accuracies*

Chapter 5:

Conclusions

Aimed at improving the accuracy of classification of speech emotions, the concept of Neural Network based, multi-level binary classification, Ensemble of classifiers was considered along with simple single stage classifier. The effect of various linear features was individually analysed Short Time Energy, Short Time Zero Cross Rate, Spectral (Roll-off, Flux and Centroid) and Pitch. The Short Time Energy based classifier offered greatest accuracy of 62.40%, but with a standard deviation of 18.77%. Although Non-Linear feature set based classification offered less accuracy of 50.40%, but the standard deviation was also least with 8.07%. The Ensemble of classifiers was constructed to improve the accuracy to 69.44%. The accuracies attained were similar to those attained in [25], [26]. Based on Ekman's six kinds of basic emotion model and Fox's multi-level emotional classification model, we have established a three level emotion recognition model which contains five classifiers for pairwise emotion classification. The model have three levels, the first level contains one classifier which divides utterance into two classes. The next two levels contain two classifiers respectively, which further refine utterances' emotions. The accuracy attained was in 68.36%, nearly equal to that attained by Ensemble of classifiers. The experiment results reveal that recognition rate of some emotions, including fear (anxiety) and disgust, still needs to be further improved.

Future Work

Further research should focus on the following aspects

- The combination of special emotional information should be paid attention to, such as the fundamental frequency rise in the end of surprising sentence, the shaking sound of fear, etc.

- Use fuzzy theory to find the probability of some kind of emotions.

- Most of the emotion research activity has been focused on advancing the emotion classification performance. In spite of the extensive research in emotion recognition, efficient speech normalization techniques that exploit the emotional state information to improve speech recognition have not been developed yet.

- Physiological patterning, in combination with facial, vocal, and other behavioural cues, can lead to significant improvements in machine recognition of user emotion, and that this recognition will be critical for giving machines the intelligence to adapt their behaviour to interact more smoothly and respectfully with people.

References (Part A)

1. E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 1997, pp. 1331–1334.
2. K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, “Speech/music discrimination for multimedia application,” in Proc. ICASSP’00, 2000.
3. J. Pinquier, J-L Rouas, R. André-Obrecht. “A Fusion Study in Speech/Music Classification”, ICASSP, pp. 17-20, 2003.
4. J. Saunders, “Real-time discrimination of broadcast speech/music,” in Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP-96), vol. 2, Atlanta, GA, May 7-10, 1996, pp. 993-996.
5. C. McKay and I. Fujinaga, “Automatic genre classification using large high-level musical feature sets,” in 5th International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain, October 2004.
6. D. Pye, “Content-based methods for the management of digital music,” in Proc. Int. Conf Acoustics, Speech, Signal Processing (ICASSP), 2000.
7. G. Tzanetakis and P. Cook, “Music genre classification of audio signals,” IEEE Trans. Speech and Audio Process., vol. 10, no. 5, pp. 293–302, Jul. 2002
8. Jiang, Dan-Ning, et al. "Music type classification by spectral contrast feature." Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on. Vol. 1. IEEE, 2002.
9. Xu, Changsheng, et al. "Musical genre classification using support vector machines." Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on. Vol. 5. IEEE, 2003.
10. Lopes, Miguel, et al. "Selection of training instances for music genre classification." Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010.
11. Bergstra, James, et al. "Aggregate features and AdaBoost for music classification." Machine learning 65.2-3 (2006): 473-484.
12. Li, Tao, and George Tzanetakis. "Factors in automatic musical genre classification of audio signals." Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.. IEEE, 2003.
13. Arnal Barbedo, Jayme Garcia, and Amauri Lopes. "Automatic genre classification of musical signals." EURASIP Journal on Advances in Signal Processing 2007.1 (2006): 1-12.
14. Li, Tao, Mitsunori Ogihara, and Qi Li. "A comparative study on content-based music genre classification." Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval.

15. McKinney, Martin F., and Jeroen Breebaart. "Features for audio and music classification." ISMIR. Vol. 3. 2003.

16. Tzanetakis, George. "Tempo extraction using beat histograms." Proceedings of the 1st Music Information Retrieval Evaluation exchange (MIREX 2005) (2005).

17. Goiiyon, Fabien. "Dance music classification: A tempo-based approach." (2004).

18. Tzanetakis, George, Georg Essl, and Perry Cook. "Audio analysis using the discrete wavelet transforms." Proc. Conf. in Acoustics and Music Theory Applications. 2001.

19. Chang, Yu-Yao, and Yao-Chung Lin. "Music Tempo (Speed) Classification." CS229 Autumn (2005).

20. Tzanetakis, George, Georg Essl, and Perry Cook. "Human perception and computer extraction of musical beat strength." Proc. DAFx. Vol. 2. 2002.

21. Lundin, Robert W. "An Objective Psychology of Music." Malabar: Robert E. Krieger Publishing Company, 1985

22. Neverman. "The Effects of Music on the Mind." 3 pp. On-line. Internet. 20 December 1999.

23. Scarantino, Barbara Anne. "Music Power Creative Living through the Joys of Music." New York: Dodd, Mead & Company, 1987.

24. Weinberger, N.M. "Threads of Music in the Tapestry of Memory." MuSICA Research Notes 4.1 (Spring 1997): 3pp.

25. Herz, R. S. "Do scents affect people's mood or work performance." Scientific American. Retrieved February 25 (2002): 2015

26. Taruskin, Richard. "Music in the Nineteenth Century: The Oxford History of Western Music." Oxford University Press, 2009

27. Laurie, Timothy (2014). "Music Genre as Method". Cultural Studies Review. 20 (2), pp. 283-292. Tagg, Philip. "Analysing Popular Music: Theory, Method and Practice". Popular Music 2 (1982).

28. Guéguen, Nicolas, and Christine Petr. "Odours and consumer behaviour in a restaurant." International Journal of Hospitality Management 25.2 (2006): 335-339.

29. Raudenbush, B., Corley, N., Eppich, W., 2001. "Enhancing athletic performance through the administration of peppermint odour." Journal of Sport & Exercises Psychology 23, 156–160.

30. Meamarbashi, Abbas, and Ali Rajabi. "The effects of peppermint on exercise performance." J Int Soc Sports Nutr 10.1 (2013): 15.

31. Sayorwan, Winai, et al. "The effects of lavender oil inhalation on emotional states, autonomic nervous system, and brain electrical activity." (2012).

32. Thoma, Myriam V., et al. "The effect of music on the human stress response." PloS one 8.8 (2013): e70156.

33. Baron, R., Bronfen, M., 1994. “A whiff of reality: empirical evidence concerning the effects of pleasant fragances on work-related behavior.” Journal of Applied Social Psychology 24, 1179–1203.

34. Diego, M., Aaron Jones, N., Field, T., Hernandez-Reif, M., Schanberg, S., Kuhn, C., McAdam, V., Galamaga, R., Galamaga, M., 1998. “Aromatherapy positively affects mood, EEG patterns of alertness and math computations.” International Journal of Neuroscience 96, 217–224.

35. Raudenbush, B., Corley, N., Eppich, W., 2001. “Enhancing athletic performance through the administration of peppermint odor.” Journal of Sport & Exercises Psychology 23, 156–160.

36. Litle, Patrick, and Marvin Zuckerman. "Sensation seeking and music preferences." Personality and individual differences 7.4 (1986): 575-578’

37. Rauscher, Frances H., Gordon L. Shaw, and Katherine N. Ky. "Listening to Mozart enhances spatial-temporal reasoning: towards a neurophysiological basis." Neuroscience letters 185.1 (1995): 44-47.

38. Altenmüller, Eckart, et al. "Hits to the left, flops to the right: different emotions during listening to music are reflected in cortical lateralisation patterns." Neuropsychologia 40.13 (2002): 2242-2256.

39. Mammarella, Nicola, Beth Fairfield, and Cesare Cornoldi. "Does music enhance cognitive performance in healthy older adults? The Vivaldi effect." Aging clinical and experimental research 19.5 (2007): 394-399.

40. Scheufele, Peter M. "Effects of progressive relaxation and classical music on measurements of attention, relaxation, and stress responses." Journal of behavioural medicine 23.2 (2000): 207-228.

41. Geethanjali, B., K. Adalarasu, and R. Rajsekaran. "Impact of music on brain function during mental task using electroencephalography." World Academy of Science, Engineering and Technology 66 (2012): 883-887.

42. Pereira, C. S., Teixeira, J., Figueiredo, P., Xavier, J., Castro, S. L., & Brattico, E. (2011), “Music and emotions in the brain: Familiarity matters.” PLoS One, 6, e27241. doi: 10.1371/journal.pone.0027241

43. Thompson WF, Schellenberg EG, Husain G. “Arousal, mood and the Mozart effect.” Psych Science 2001; 12: 248-51.

44. Sakharov D S, Davydov V I, Pavlygina R A. “Inter Central Relations of the Human EEG during Listening to Music”. J Human Physiology 2005 (31), pp: 27 – 32

45. Koelsch S A. “Neural basis of music-evoked emotions. Trends in Cognitive Sciences” 2010 (14), pp: 131 – 137

46. Peretz I, Zatorre R. “Brain Organization for Music Processing. Annual Review of Psychology” 2005 (56), pp: 89 – 114

47. Lonsdale, A. J., & North, A. C.. Why do we listen to music? A uses and gratifications analysis. British journal of psychology London England 1953, 2011 pp: 102 (1), 108-134.

48. Kirschbaum, Clemens, K-M. Pirke, and Dirk H. Hellhammer. "The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting." Neuropsychobiology 28.1-2 (1993): 76-81.

49. Manly, T., Owen, A.M., McAvinue, L., Datta, A., Lewis, G.H., Scott, S.K., Rorden, C., Pickard, J., Robertson, I.H. (2003). "Enhancing the sensitivity of a sustained attention task to frontal damage: convergent clinical and functional imaging evidence." Neurocase, 9(4), 340-349

50. Scheibe, K. E., Shaver, P.R. and Carrier, S.C.(1967). "Colour association values and response interference on variants of the Stroop test." Acta Psychologica, 26: 286-95.

51. Scheibe, K. E., Shaver, P.R. and Carrier, S.C.(1967). "Colour association values and response interference on variants of the Stroop test." Acta Psychologica, 26: 286-95.

52. Gfeller, Kate. "Musical components and styles preferred by young adults for aerobic fitness activities." Journal of Music Therapy 25.1 (1988): 28-43.

53. Karageorghis, Costas I., et al. "Psychophysical and ergogenic effects of synchronous music during treadmill walking." (2009).

54. Diego, M., Aaron Jones, N., Field, T., Hernandez-Reif, M., Schanberg, S., Kuhn, C., McAdam, V., Galamaga, R., Galamaga, M., 1998. "Aromatherapy positively affects mood, EEG patterns of alertness and math computations." International Journal of Neuroscience 96, 217–224.

55. R. N. Shepard, "Circularity in judgements of relative pitch," J. Acoust. Soc. Amer., vol. 36, pp. 2346–2353, 1964.

56. Bartsch, Mark A., and Gregory H. Wakefield. "Audio thumb nailing of popular music using chroma-based representations." Multimedia, IEEE Transactions on 7.1 (2005): 96-104.

57. Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." Speech and Audio Processing, IEEE transactions on 10.5 (2002): 293-302.

58. Tagg, Philip. "Analysing Popular Music: Theory, Method and Practice". Popular Music 2 (1982): 41.

59. Diego MA, Jones NA, Field T, Hernandez-Reif M, Schanberg S, Kuhn C, et al. "Aromatherapy positively affects mood, EEG patterns of alertness and math computations." Int J Neurosci 1998; 96: 217-24.

60. Mayer, John D. "Brief mood introspection scale (BMIS)." (2008).

References (Part B)

1. Cabanac, Michel (2002). "What is emotion?" Behavioural Processes 60(2): 69-83.
2. Schacter, Daniel L. (2011). Psychology Second Edition. 41 Madison Avenue, New York, NY 10010: Worth Publishers. p. 310. ISBN 978-1-4292-3719-2.
3. Plutchik, Robert, and Henry Kellerman, eds. “Theories of emotion.” Vol. 1. Academic Press, 2013
4. Gaulin, Steven J. C. and Donald H. McBurney. Evolutionary Psychology. Prentice Hall. 2003. ISBN 978-0-13-111529-3, Chapter 6, p 121-142.
5. Schwarz, N. H. (1990). “Feelings as information: Informational and motivational functions of affective states.” Handbook of motivation and cognition: Foundations of social behaviour, 2, 527-561.
6. Handel, Steven. "Classification of emotions." (2012).
7. Dimitrios Ververidis, Constantine Kotropoulos, “Emotional speech recognition: Resources, features, and methods”, Speech Communication 48 (2006) 1162–1181.
8. T. Polzin: "Verbal and non-verbal cues in the communication of emotions", ICASSP 2000, Paper Proc. ID 3485, Turkey, 2000.
9. R. Picard, E. Vyzas, and J. Healey, “Toward machine emotional intelligence,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1175–1191, 2001.
10. K. S. Rao, R. Reddy, S. Maity, and S. G. Koolagudi, “Characterization of emotions using the dynamics of prosodic features,” in International Conference on Speech Prosody, (Chicago, USA), May 2010.
11. Shashidhar G. Koolagudi, Ramu Reddy and K. Sreenivasa Rao, “Emotion recognition from speech signal using epoch parameters”, Signal Processing and Communications - SPCOM, 2010 Int. Conf., ISBN: 978-1-4244-7137-9, pp. 1-5
12. D. Ververidis, C. Kotropoulos, I. Pitas, “Automatic emotional speech classification”, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, Proceedings, (ICASSP ’04), vol. 1, 2004, pp. I-593-6.
13. L. J. Chen, X. Mao and Y. L. Xue, “Speech emotion recognition: Features and classification models”, Digital signal processing, vol.22, pp. 1154-1160, 2012.
14. S. G. Kooladugi, N. Kumar, and K. S. Rao, “Speech emotion recognition using segmental level prosodic analysis,” in Proc. Int. Conf. Devices and Communications, 2011, pp. 1–5.
15. C. Busso, S. Lee, and S. Narayanan, “Using neutral speech models for emotional speech analysis,” in Proc. Interspeech’07—Eurospeech, Antwerp, Belgium, Aug. 2007, pp. 2225–2228.

16. L. R. Rabiner and R.W. Schafer, “Digital Processing of Speech Signals”.

17. L. R. Rabiner, M. J. Vheng, A.E. Rosenberg, and C.A. McGonegal, “A Comparitive performance study of several pitch detection algorithms,” IEEE Trans. Acoust. Speech, and Signal Proc., Vol. ASSP-24, No. 5, pp. 399-418, 1976.

18. B. Gold, “Computer Program for Pitch Extraction,” J. Acoust. Soc. Am. Vol. 34, No.7, pp. 916-921, 1962

19. M. J. Ross, H.L. Shaffer, A. Cohen, R.Freudberg, and H.J. Manly, “ Average Magnitude Difference Function Pitch Estimator,” IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-22, pp. 353-362, October 1974

20. Cairns, D., Hansen, J.H.L., 1994, “Nonlinear analysis and detection of speech under stressed conditions.” J. Acoust. Soc. Am. 96 (6), 3392–3400.

21. Kaiser, “On a simple algorithm to calculate the “energy” of a signal,” in Proc. Int. Conf. Acoustic, Speech, Signal Processing ’90, 1990, pp. 381–384.

22. Kaiser, “On Teager’s energy algorithm, its generalization to continuous signals,” in Proc. 4th IEEE Digital Signal Processing Workshop. New Paltz, NY, Sept. 1990.

23. G. Zhou, John H. L. Hansen, “Nonlinear Feature Based Classification of Speech Under Stress”, IEEE transactions on speech and audio processing, vol. 9, no. 3, March 2001

24. Dietterich, T.G. (2001): “Ensemble methods in machine learning. In Kittler, J., Roli, eds.: Multiple Classifier Systems.” LNCS Vol. 1857, Springer (2001) 1–15.

25. Schuller, B., Rigoll, G., 2006. “Timing levels in segment-based speech emotion recognition.” In: Proc. Interspeech, Pittsburgh, PA, USA, pp. 1818–1821.

26. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A., 2009e. “Acoustic emotion recognition: a benchmark comparison of performances.” In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Merano, pp. 552–557.

27. Teager, H. M., and S. M. Teager. "Active fluid dynamic voice production models, or there is a unicorn in the garden." Vocal fold physiology: biomechanics, acoustics, and phonatory control (1983): 387-401.

28. Teager, Herbert M. "Some observations on oral air flow during phonation." Acoustics, Speech and Signal Processing, IEEE Trans on 28.5 (1980): 599-601.

29. Teager, H. M., and S. M. Teager. "The effects of separated air flow on vocalization." ed. IR Titze & RC Scherer, Vocal Fold Physiology (1983): 124-141.

30. Burkhardt, Felix, et al. "A database of German emotional speech." Interspeech. Vol. 5. 2005.

31. Ekman, Paul. "An argument for basic emotions." Cognition & emotion 6.3-4 (1992): 169-200.

Appendixes

Appendix A

Consent to participate in effect of Aroma on physical performance, mental performance and relaxation of a subject.

Follow-up Informed Consent Form for Use of Data in Experiment.

☐ Trier Social Stress Test

☐ Mental Performance Test

☐ Physical Performance Test

I _____ have just provided informed consent and participated as a subject in an experiment to study the effect of aroma on physical, mental performance and relaxation effect.

At this point, I understand that if I object to having undergone this procedure without my knowledge, I may choose to withdraw my data from the study with no prejudice to me and no loss of subject pay.

Please sign the following statement to give informed consent:

I agree to have my data used further in this experiment_____

Date:

Please fill the follow-up table for further analysis:

Name:	
Age:	
Sex (M/F):	
Weight (in kgs):	
Height (in cms):	
Any serious Health Issue (Y/N):	
Any Experience in Aromatherapy (Y/N):	
Any Extensive experience in Music (Y/N):	
Any Drug Addiction (Y/N):	

Appendix B

Aroma Release Systems

Aromatherapy (diffusing essential oils) helps to stimulate the nervous system and provides a host of health benefits including creating a stress-free environment, relieving tension, providing energy, improving mental clarity, clearing nasal passages and reducing bacteria, fungus from your home. Diffusing essential oils disperses the oils into the air filling the room with a wonderful scent. Using essential oils for aroma is non-toxic with the added benefits of being therapeutic and healthy for you. There are four basic methods of releasing Essential Oils:

- **Nebulizing Diffusers:** Nebulizing diffusers are often considered the most powerful type of diffusers and with good reason. They do not need water or heat to get the essential oil in to the air and they work by using an atomizer to create fine, airborne particles of essential oils and blowing them in to the air. A 15 minute working will leave the aroma for hours.
- **Ultrasonic Diffusers:** Ultrasonic diffusers work in a similar way to nebulizing diffusers by creating a fine mist. The difference is that ultrasonic diffusers use water and essential oils to create a cool mist of water/oils that releases in to the air. They double as a humidifier, so they are beneficial in winter, but they don't put out as strong of a concentration of essential oils since they also use water.
- **Heat Diffuser:** The heat diffuser uses heat to diffuse essential oil. It affects both the healing benefits for oil and the intensity of the fragrance, as it heats the essential oil. Some of these diffusers use high levels of heat to produce an extra strong aroma that can change the chemical bonding of the essential oils. Therefore, a diffuser that uses very low levels of heat is considered better.
- **Evaporative Diffuser:** diffusers scatter essential oils with a fan blowing method with an attached filter. Evaporative diffusers do not disperse the complete blend of essential oils at one time. They tend to portion out the oils. Lighter components are evaporated quicker than heavier components. This apportioning of the components (oils/water) cuts down on the overall therapeutic value of the essential oils. Another drawback of these diffusers is that they have shorter running times.